Detecting Factors Responsible for Diabetes Prevalence in Nigeria using Social Media and Machine Learning

Oladapo Oyebode
Faculty of Computer Science
Dalhousie University
Halifax, Canada
oladapo.oyebode@dal.ca

Rita Orji
Faculty of Computer Science
Dalhousie University
Halifax, Canada
rita.orji@dal.ca

***Abstract*—Diabetes is a non-communicable disease associated with increased level of glucose due to inadequate supply of insulin (known as Type 1 diabetes) or inability to use insulin efficiently (known as Type 2 diabetes). Though the exact cause of Type 1 diabetes is unknown, the probable causes are genetics and environmental factors (such as exposure to viruses). On the other hand, Type 2 diabetes is largely linked to unhealthy lifestyle choices. In Nigeria, many people are believed to be living with diabetes and the country's diabetes prevalence rate is one of the highest in Africa. To determine the factors responsible for diabetes prevalence in Nigeria, we analyzed social media contents related to diabetes since billions of people, including diabetic patients and healthcare professionals, use social media platforms to freely share their experiences and discuss many health-related topics. None of the existing research targets the African audience who are also major users of social media platforms; hence our work aims to close this gap by leveraging an African social media platform targeted at Nigerians to gather diabetes-related data, and then applying machine learning technique to detect those factors responsible for diabetes prevalence in Nigeria. Based on our results, we discussed positive behavioural or lifestyle changes that are necessary to prevent and treat diabetes in Nigeria, as well as intervention designs required to bring about those changes. Future work will develop a diabetes intervention application implementing all the design features highlighted in Section V of this paper and making it generally accessible to Nigerians.**

Keywords—Diabetes, Prevalence factors, Social media, Text mining, Machine learning, Intervention design

I. INTRODUCTION

Social media provides an opportunity to reach most internet users with health-related information. Health authorities and promoters, as well as healthcare professionals, now communicate with target audiences on social media platforms (such as Facebook, Twitter, and others) due to their large user base (over 2 billion active users [12]) at little or no cost. These platforms also allow people with internet access to share information and respond to health-related topics. Interestingly, Nigeria ranks 8th among top 20 countries with highest number of internet users in the world [6]. Over 98 million Nigerians are active internet users [7] and over 75% of them are active on social media [11]. Apart from Facebook, Twitter, and Instagram, there are other social media platforms targeted at specific users or users from specific geographic locations, one of which is Nairaland. Nairaland is an African online community launched in 2005, and it is targeted at Nigerians. Nairaland is used by over 55 million Nigerians [16], has 38 forums generating over 4 million topics [8], and over 219 million page views [16]. Examples of Nairaland forums include Health forum, Travel forum, Business forum, Politics forum, Agriculture forum, and so on. Each forum has multiple topics, with each topic focusing on a specific life event or issue so people can send comments in form of posts. Our interest is in the Health forum, especially topics related to diabetes. We focused on diabetes because people diagnosed with non-communicable diseases (NCDs) are on the rise in Nigeria, and deaths from NCDs have been estimated to increase by 27% in Africa by 2030 [9]. For instance, 11.2 million people are believed to be living with diabetes in Nigeria [13]. In other words, the prevalence rate of diabetes in Nigeria with a total population of 193.3 million [13] is 5.77%. Our work involves mining health-related topics and all the posts under each topic from Nairaland using the web scraping technique, applying a filtering mechanism to determine posts related to diabetes, classifying our dataset into posts addressing diabetes causatives (and those that are not) using machine learning technique, and finally visualizing and interpreting our results.

The main contribution of this work is to detect factors responsible for diabetes prevalence in Nigeria, and then suggest some design features that could be employed to design effective diabetes technological intervention targeted at Nigerians. As a result, Nigerians are empowered to take charge of their health through lifestyle changes that are measurable and effective, thereby constituting a viable approach for curtailing the prevalence of diabetes.

II. RELATED WORK

Since the advent and prevalence of social media, research has begun to investigate how it could be used to promote public and social good, such as health promotion. As a result, several research has investigated how people use the social media for health-related purposes with the aim of informing social media-based health intervention design. For example, Rani et al. [10] extracted health tweets from Twitter, and applied Naïve Bayes model to classify tweets into those containing "diabetes" keyword and those without. Afterwards, they extracted features

978-3-903176-24-9

with their contribution counts using Support Vector Machine (SVM) and then applied their Highest-Ranking Feature Extraction Algorithm to identify features contributing more to diabetes. They combined these features with the "Diabetes" keyword to form bigrams (such as Diabetes-Insulin, Diabetes-Type1, and so on) upon which they applied Term Tweet Classification Matrix to determine their strengths and contributions towards developing an efficient diabetes tracking tool.

Lincke et al. [4] investigated the content and flow of information whenever the Swedish people tweets about diabetes-related issues. They extracted diabetes tweets written in Swedish using keywords, such as "diabetes", "blodsocker", "insulin", "hypogly", "HbA1c", "flashmatar", and "flash-matar". Having performed necessary clean-up by removing duplicates, penetrations, stop words, and applying stemming on words, they applied k-means clustering (an unsupervised machine learning technique) to group similar discussion topics.

Furthermore, Hasan et al. [2] built two machine learning classification models using Naïve Bayes and Support Vector Machines (SVM) to compare three sentiment analyzers (W-WSD, TextBlob, and SentiWordNet). Naïve Bayes produced better accuracy than SVM.

Other related works by Wong et al. [17] and Wagh et al. [15] have applied machine learning classification techniques on social data.

III. Methodology

We used text mining and machine learning techniques to explore, process, and classify our dataset such that we can determine the factors responsible for diabetes prevalence in Nigeria. Although most of the posts are based on individual experiences or knowledge and some by healthcare professionals, our approach can expose new insights that may be useful for healthcare providers, Nigerians and Africans in general with respect to informing diabetes-intervention design.

A. Data Collection

From Nairaland, we extracted 371,996 posts from 74,224 topics within the Health forum using the web scraping technique since Nairaland does not have an application programming interface (API). We implemented the web scraper using *rvest* package of R language [5]. In order to select only topics related to diabetes, we applied a filtering mechanism which extracts only topics containing any of the following keywords: *diabetes*, *diabetics*, *diabetis*, *diabeties*, *diabetic*, *diabetese*, and *diabete*. After applying this filtering technique on the topics, we ended up with 3,051 posts from 872 topics. The 3,051 diabetes-related posts become our dataset or corpus which is then preprocessed and analyzed.

B. Data Preprocessing

After data extraction, we performed the following steps using the text mining (tm) framework [1] available through the *tm* package in R, with the aim of preparing our data for classification using machine learning:

- Remove punctuation and special characters
- Convert texts to lowercase
- Remove numbers
- Remove stop words that do not add value, such as "a", "an", "the", "has", "him", "can", "also", "will", "get", and so on
- Remove extra whitespaces
- *Create bag of words:* This process involves breaking texts into terms (or unigrams) by creating a document term matrix (DTM) in which rows correspond to documents (or posts) and columns correspond to terms. Each element of the DTM represent the term frequency (i.e. number of times a term or word occurs in a document). **The DTM becomes the main dataset for our classification task**.
- *Select features or terms that are useful from the DTM and ignore those that are not*: To achieve this, we identified infrequent terms and then restricted the DTM to use only frequent terms whose total frequency count across all posts or documents is at least 80.
- Remove words with high frequency count but not contributing to the overall objective

C. Data Classification

First, we manually labelled the posts in our dataset into two classes - the "causative" class and "other" class for training purpose. The "causative" class means corresponding post explains one or more factors that cause diabetes, while the "other" class explains other unrelated issues.

Next, we handled the imbalance in our dataset by applying the random under-sampling technique [3] to randomly select 30% of the majority class ("other" class) such that the majority to minority class ratio becomes 794:406 rather than the initial 2,645:406. We retained some imbalance, though minimal, to avoid losing much vital information that can aid better prediction.

Afterwards, we partitioned our dataset into 80% training set and 20% test set. Our training set is used to train a Naïve Bayes machine learning model. We chose Naïve Bayes machine learning algorithm for our classification task because it is efficient and can handle large number of features. The test set is used for model evaluation.

1) Developing Our Model: Naïve Bayes algorithm has three variants – the Multinomial, Binarized, and Bernoulli Naïve Bayes [14]. We selected the Binarized Naïve Bayes (BNB) algorithm since it places significance on word occurrence rather than word frequency. To apply the BNB algorithm, we transformed both the training set and test set such that every word frequency greater than 0 is set to "1", and every word frequency not greater than 0 is set to "0". Afterwards, we built our model using the *e1071* package in R which provides the *naiveBayes* function that accepts our training set, the corresponding class labels, and the LaPlace smoothing as arguments.

IV. Results

We used our model to predict the class of each post in our test set. The model accuracy is given by the proportion of total

number of correct predictions. As shown in Fig. 1, the accuracy of the model is 87.08%, which means our model did very well at predicting correct class for each post. Based on the confusion matrix in Fig. 2, our model was able to predict correct classes for 209 out of 240 posts in the test set. Furthermore, the precision and recall metrics explain how well our model does in predicting our class of interest (the "causative" class) correctly. With values 79.52% and 82.5% for precision and recall respectively (see Fig. 1), our model performed well, considering the minimal imbalance in our dataset. The F1 score of 80.98% is the harmonic mean of precision and recall.

In determining the factors responsible for diabetes prevalence, we analyzed the unigrams for the posts classified by the model as "causative". As shown in Fig. 3, our results revealed the following facts:

- Weight, Food, and Diet are the major factors responsible for diabetes prevalence in Nigeria.
- Other factors, though less prominent, contributing to diabetes include pregnancy, age, sleep, and others

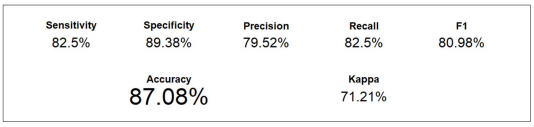

Fig. 1. Model Performance

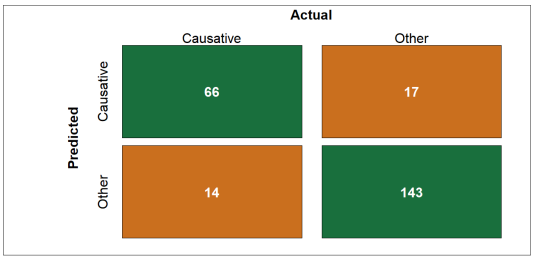

Fig. 2. Confusion Matrix

Fig. 3. Wordcloud showing unigrams and their prominence

Our results are therefore in line with known facts about Type 2 diabetes.

V. Discussion

It is not surprising that Weight, Food, and Diet are predominant diabetes causatives, considering the Nigerian culture and environment. For instance, most Nigerian foods are rich in carbohydrate/starch and fat; hence the possibility of increased blood glucose level and weight gain is high. Also, most Nigerians do not have the culture of exercising and with increasing industrialization, most people practice sedentary lifestyle due to the nature of their jobs. Traffic situations, caused by increasing population and heavy dependence on road transport, in major urban cities in Nigeria also contribute immensely to sedentary lifestyle since people spend hours sitting in cars and buses before reaching their destinations. This sedentary lifestyle will lead to weight gain if constant physical activity or exercise is missing. Treatment for Type 2 diabetes without medication can be achieved largely through lifestyle changes. Thus, effective diabetes prevention and natural treatment interventions which promote healthy behaviours leading to weight reduction, increased physical activity, selecting healthier foods and maintaining a healthy diet among Nigerians will suffice.

For technology-savvy Nigerians, mobile (or web) applications for monitoring diabetes triggers or treating diabetes will be appealing. For an application to be fit for this purpose, it must possess the features that entrench these healthy behaviours in users. Table I summarizes the recommended features that can bring about the healthy behaviours. The first two features are based on the three key and most prominent factors in our results, while the last two features are based on the less prominent factors.

Another important tool for sharing diabetes-related information and accelerating behaviour change is the social media. Diabetes prevention and treatment knowledge-sharing sessions, targeted at Nigerians and African audience in general, can be launched by healthcare professionals on web and mobile social media platforms used frequently by the audience, such as Facebook, Twitter, Instagram, WhatsApp, and so on. The sessions will involve sharing information related to weight reduction strategies, foods and diet for self-treating diabetes, physical activity or exercise techniques, and other related information on a "dedicated" page, handle, account, or group on each of these platforms. The benefit of such sessions is that the target audience are able to access accurate information anytime and anywhere, and use these information to take control of their lifestyle by setting goals, such as weight reduction, diet, and exercise goals, and monitoring their progress. Another benefit is that questions, concerns, and progress reports are easily communicated to healthcare professionals directly anytime, thereby promoting unhindered interaction between healthcare professionals and the audience.

TABLE I. RECOMMENDED FEATURES IN DIABETES PREVENTION OR TREATMENT APPLICATIONS

Feature	Description
Weight and Physical Activity tracking	Allow users to set weight loss and exercise goals, and then track progress made towards achieving those goals. Constant reminder is needed to encourage users to persist.
Diet monitoring	Suggest (and allow users to search for) healthier diets that lower risk of diabetes. For instance, a better diet may include unprocessed foods, much vegetables, fish, low-fat dairy products, fruits, whole-grains, and so on.
Stress management	Educate on stress management strategies since stress reduction leads to lower blood pressure as well. Lower blood pressure reduces the risk of diabetes.
Sleep tracking	Track sleep and educate on achieving healthy sleep levels.

For those without smartphones and internet access, the Nigerian government in conjunction with healthcare professionals and health promoters could launch a comprehensive lifestyle program (available at selected venues across the country) with trainings during each session, covering weight loss, healthy diet, and physical activity. There should be weight, diet and exercise goals (which can be either daily, weekly, bi-weekly, or monthly goals) that each eligible participant is expected to meet. To track progress, participants will be required to weigh themselves at each session, record their exercise, as well as what they eat. The lifestyle program can last for a year or less, enough to empower participants to take charge of their health and effectively prevent and control diabetes.

VI. CONCLUSION AND FUTURE WORK

The current prevalence rate of diabetes in Nigeria is high and can double in the coming years if drastic actions are not taken quickly. Our research work applied machine learning technique on health-related posts (mined from Nairaland, a widely used social media platform targeted at Nigerians) to determine the factors responsible for diabetes prevalence in Nigeria. These factors will guide health intervention designers, health promoters, as well as the Nigerian government through the Federal Ministry of Health (FMOH) and the Diabetes Association of Nigeria (DAN), and healthcare professionals in their efforts towards curbing diabetes in the country. Our work further explains how to empower individuals to take charge of their health through lifestyle changes that are measurable and effective, thereby constituting a viable approach for curtailing the prevalence of diabetes in Nigeria. We suggest some design features that could be employed to design effective diabetes technological intervention targeted at Nigerians.

The next phase of this work is to develop a diabetes intervention application implementing all the features highlighted in Table I and making it generally accessible to Nigerians.

REFERENCES

[1] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R," *J. Stat. Softw.*, vol. 25, no. 5, pp. 1–54, 2008.

[2] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," *Mathematical and Computational Applications*, vol. 23, no. 1**,** 2018.

[3] C. Kuo, "Using Under-Sampling Techniques for Extremely Imbalanced Data." [Online]. Available: https://medium.com/anomaly-detection-with-python-and-r/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8. [Accessed: 18-Nov-2018].

[4] A. Lincke, J. Lundberg, M. Thunander, M. Milrad, J. Lundberg, and I. Jusufi, "Diabetes Information on Social Media," in *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction*, 2018.

[5] W. Marble, "Web Scraping With R," 2016. [Online]. Available: http://stanford.edu/~wpmarble/webscraping_tutorial/webscraping_tutorial.pdf. [Accessed: 24-Sept-2018].

[6] Miniwatts Marketing Group, "Top 20 Countries With The Highest Number Of Internet Users." [Online]. Available: https://www.internetworldstats.com/top20.htm. [Accessed: 24-Sept-2018].

[7] Miniwatts Marketing Group, "Internet Usage in Africa." [Online]. Available: https://www.internetworldstats.com/africa.htm. [Accessed: 24-Sept-2018].

[8] Nairaland, "Nairaland Stats." [Online]. Available: https://www.nairaland.com/. [Accessed: 24-Sept-2018].

[9] Nigeria Health Watch, "The weighty burden of NCDs in Nigeria: Time to Act," 2018. [Online]. Available: https://medium.com/@nigeriahealthwatch/the-weighty-burden-of-ncds-in-nigeria-time-to-act-9a47fc4a3a27. [Accessed: 31-Dec-2018].

[10] V. V. Rani, and K. S. Rani, "Efficient Tool for Diabetes Tracking through Layered Bigram Approach*," IADS International Conference on Computing, Communications & Data Engineering (CCODE)*, 2018.

[11] StatCounter, "Social Media Stats Nigeria." [Online]. Available: http://gs.statcounter.com/social-media-stats/all/nigeria. [Accessed: 31-Dec-2018].

[12] Statista, "Number of social network users worldwide from 2010 to 2021 (in billions)." [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. [Accessed: 31-Dec-2018].

[13] A. E. Uloko *et al.*, "Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis," *Diabetes therapy: research, treatment and education of diabetes and related disorders*, vol. 9, no. 3, pp. 1307–1316, 2018.

[14] V. Vryniotis, "Machine Learning Tutorial: The Naive Bayes Text Classifier," 2013. [Online]. Available: http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/. [Accessed: 24-Sept-2018].

[15] B. Wagh, J. Shinde, and P. Kale, "A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques," *International Journal of Emerging Research in Management and Technology*, 2018.

[16] Wikipedia, "Nairaland," 2018. [Online]. Available: https://en.wikipedia.org/wiki/Nairaland. [Accessed: 24-Sept-2018].

[17] S. Wong, H. Chinaei, and F. Rudzicz, "Predicting health inspection results from online restaurant reviews," 2016. [Online]. Available: https://arxiv.org/pdf/1603.05673.pdf. [Accessed: 24-Sept-2018].