ANALYZING ONLINE MEDIA ARTICLES ON DIABETES USING NATURAL LANGUAGE PROCESSING: A COMPARATIVE STUDY OF INDIAN OCEAN REGION AND FRANCE

Mohammud Shaad Ally Toofanee[1,2], Nabeelah Zainab Ally Pooloo[2], Sabeena Dowlut[2], Karim Tamine[1], and Damien Sauveron[1]

[1]XLIM, UMR CNRS 7252, University of Limoges, 123,Avenue Albert Thomas, 87060 Limoges, France
[2]Universit´e des Mascareignes, Concorde Avenue Roches Brunes Rose Hill, Mauritius

ABSTRACT

***Background**: Diabetes is a global health concern affecting millions of people worldwide. However, knowledge, attitudes, and practices related to this disease vary widely across different regions. This article aims to investigate mediainfluenced perceptions about diabetes in France and the Indian Ocean countries using natural language processing (NLP) techniques applied to online news articles. Findings aims to provide expert in Health Literacy (HL) and health promotion to develop better communication strategies. **Method**: Constitute a datatset of Online news articles on Diabetes and apply NLP like Word2Vec for word integration, LDA for topic identification, and transformer-based classification models (e.g., BERT and its variants) for sentiment analysis. processing (NLP). **Results**: Sentiment analysis revealed more negative discussions about diabetes in the Indian Ocean region (48%) compared to France (32%), with neutral articles dominating in France (42%). In terms of topic Identification there were some topic which appeared for France which were not present for indian ocean region. **Discussions**: The findings of this study indicate that perceptions and discussions about diabetes differ between two regions, which have implications for public health interventions and communication strategies. However, the study is limited by the initial amount of information captured for analysis.*

KEYWORDS

Artificial Intelligence, Natural Language Processing, Mass Media, Diabetes, LDA,Transformers, BERT, Sentiment Analysis, Word Associations

1. INTRODUCTION

The internet has become the predominant source of information for individuals when encountering various problems or health conditions. A recent study concluded that 74.4% of individuals in the United States initially sought healthrelated information online, while only 13.3% consulted a healthcare professional as their first step[1]. However, online platforms often contain poorly reported information, including misinformation, cherry-picked data, exaggerated claims, and other misleading content, posing a significant risk to public health [2]. This issue has been widely observed across different regions of the world, particularly in the context of the

David C. Wyld et al. (Eds): NLPML, AIAP, SIGL, CRIS, COSIT, DMA -2023
pp. 199-212, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130817

COVID-19 pandemic. Additionally, previous research has highlighted the role of mass media communication in shaping public opinions[3]

Type 2 Diabetes (T2D) is a significant health concern in Mauritius, with the International Diabetes Federation (IDF) predicting a prevalence rate of 26.6% by the year 2045 [4]. In this study, we propose utilizing artificial intelligence methods, specifically natural language processing (NLP), to analyze online media articles discussing T2D in countries within the region and compare them with a developing country in Europe, namely France. Our objective is to provide health professionals with some additional inputs to adapt communication message and strategies for T2D prevention and education in these regions.

NLP enables the exploitation and manipulation of a vast amount of data to gain insight that would otherwise not be humanly possible. However, NLP tends to confirm clinical hypotheses rather than develop entirely new information [5]. while several researches has been carried out on information exchanged on social media platforms, mainly twitter, this is not possible for the local context. In the local context talking about a pathology remains taboo hence the important to analyse media topics and sentiments since it is one of the main source of information on health related issues and the impact in has on public opinion The dataset constituted from this research is also intended to be used as input for a project on AI powered chabot for diabetes prevention and management.

Natural Language Processing (NLP) as a powerful tool for extracting and analyzing large amounts of data to gain insights that may not be easily attainable through human efforts. While many studies have utilized social media platforms, such as Twitter, for health-related information analysis, this approach may not be feasible in local contexts where discussing certain health issues is considered taboo. Thus, it is essential to analyze media topics and sentiments as they serve as a valuable source of information regarding health issues and their influence on public opinion.

Based on both psychology and sociology, the framing effect theory explains the ability of news media to affect people’s attitudes and behaviors through making slight changes[6] The findings of this comparative study can provide valuable insights for designing effective communication strategies and interventions to address the complex social and psychological dynamics associated with diabetes.

This work focuses on three key text mining tasks: sentiment analysis, automated topic extraction (Topic Modeling), and semantic correlation of words with their context related to diabetes in online news media. The dataset generated from this study is intended to be used as part of the training for an AI-powered chatbot project aimed at diabetes prevention and management.

2. Related Works

Deep learning has being widely utilized in the medical field, particularly in the analysis of medical digital images. However, there has been a recent trend towards exploring the potential of textual data as well. In a recent study, Boissonnet et al. (2002) proposed a model for evaluating the quality of health-related articles and providing explanations for the classification they make, utilizing the BERT (Bidirectional Encoder Representations from Transformers) approach [2]. This work contributes to the growing body of research on leveraging NLP techniques for analyzing textual data in the context of health-related articles.

In their recent publication, UnKyo et al. (2002) conducted a study on sentiment analysis of telemedicine-related newspaper articles during the COVID-19 pandemic in Korea, investigating the association between the pandemic and changes in the media's perception of telemedicine. They employed Latent Dirichlet allocation (LDA) analysis, topic extraction, and topic trend analysis to analyze the data and draw conclusions [6]. This research contributes to the understanding of how the perception of telemedicine has evolved during the pandemic. Furthermore, in a separate study, Wang et al. (2022) aimed to demonstrate the applicability of natural language processing (NLP) techniques in the Chinese language medical environment. They successfully utilized NLP techniques to rapidly identify vulnerability factors in the management of Type 2 diabetes (T2DM) [7]. Their research highlights the potential of NLP in uncovering insights in the Chinese language medical domain.

Int the study conducted by Oyebode et al. (2019), they applied natural language processing (NLP) techniques to social media data collected from Nigerian platforms to detect factors responsible for diabetes prevalence [8]. Using the Binarized Na¨ıve Bayes (BNB) algorithm, their solution revealed significant factors such as weight, diet, pregnancy, age, and sleep that contribute to diabetes prevalence. These findings provide valuable insights for actors in the health sector to guide interventions in diabetes education and prevention efforts.Building upon their previous work, the same authors developed a medical named entity recognition framework called MediNER, which effectively identifies named entities related to diabetes management and classifies them into categories such as Food, Medication, Therapeutic Procedure, and Supplement [9]. This research showcases the potential of NLP techniques in improving diabetes management through precise identification of relevant entities.

The mining of data exchanged on social media platforms regarding COVID-19 and vaccination has also gained significant attention from the research community, with studies conducted by Praveen et al. (2022), Canaparo et al. (2023), and Zulfiker et al. (2022)[10,11,12].

Additionally, several researchers have applied NLP techniques for sentiment analysis related to diabetes and social networks, as demonstrated in the works of Gabarron et al. (2019), Salas et al. (2017), De et al. (2012), and Liu et al. (2020) [13,14,15,16]. These studies collectively highlight the increasing interest in utilizing NLP techniques to gain insights from social media data in the context of diabetes and related health issues.

Foley et al.(2020) researched how diabetes pandemic was portrayed in the United kingdom news bewteen 1993-2013[17] and furthermore Syafhan et al. investigated on media reporting of antidiabetic medicines in newspapers published in the United Kindgom and United states[18]. Such endeavors hold immense potential for NLP experts to efficiently analyze vast amounts of information and generate numerous insights, thus making it a crucial area of research.

3. Materials and Methods

The figure 1 presents an overview of the tasks that need to be completed to achieve the objectives initially set with the first step being data collection. Each steps mentioned are explained in details in the following sections.

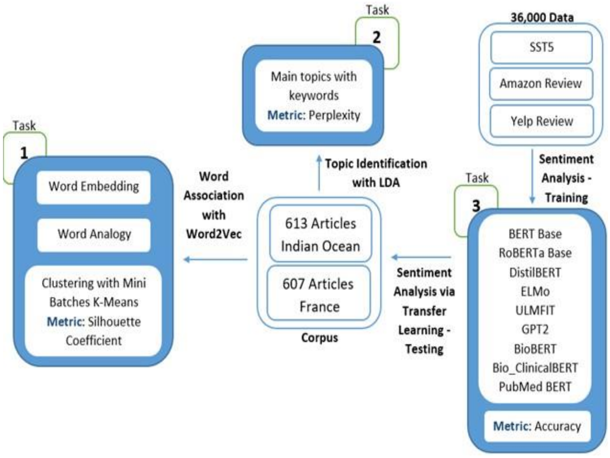

Task
1
Word Embedding
Word Analogy
Clustering with Mini Batches K-Means
Metric: Silhouette Coefficient
Word Association with Word2Vec
Task
2
Main topics with keywords
Metric: Perplexity
Topic Identification with LDA
613 Articles Indian Ocean
607 Articles France
Corpus
36,000 Data
SST5
Amazon Review
Yelp Review
Sentiment Analysis - Training
Task
3
BERT Base
RoBERTa Base
DistilBERT
ELMo
ULMFIT
GPT2
BioBERT
Bio_ClinicalBERT
PubMed BERT
Metric: Accuracy
Sentiment Analysis via Transfer Learning - Testing

Fig.1: Overview of Task to be completed

3.1. Dataset and Pre-processing

The first step involved collecting textual data from online newspaper publications in two regions, the Indian Ocean and France, in order to analyze specificities related to Type 2 Diabetes. We used web scraping techniques, such as Beautiful Soup, ParseHub[19], and Octoparse[20], to extract electronic text data from structured web pages of news outlets and health magazines focused on diabetes. The collected data was consolidated and filtered to include only articles that were less than five years old to ensure relevance to current lifestyle practices and diabetes conditions. Pre-processing techniques were then applied to prepare the textual data for analysis.

1. Removing duplicate cells, empty values, punctuation marks, URLs,
2. Tokenisation, which is the process of dividing the text into smaller units,
3. Stemming, is a natural language processing technique that lowers inflectionin words to their root forms,
4. Lemmatisation, which is the process of ensuring that etymological words donot lose their meaning.

3.1. Vectorisation, Word Association and Word Analogy

After data collection and pre-processing, Word2Vect method was applied for vectorisation.Word2vec is a technique of natural language processing [21] that is used to produce word embeddings. Word2vec takes as its input a large corpus of text and produces a vector representation, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space [21]].

Wwe utilized the Word2Vec algorithm to implement word embedding, and performed word analogy analysis to gain insights into word associations. Popular libraries such as Keras and Gensim were employed for these tasks, with a focus on Gensim for its ability to visualize word embeddings and identify similar words. Additionally, the performance of the word embedding was evaluated through vector arithmetic-based word analogy examinations. The findings of this study shed light on the semantic relationships between words in the context of diabetes, providing valuable insights for further research in the field.

3.2. Latent Dirichlet Allocation (LDA) Analysis , Topic Identification and Modeling

Latent Dirichlet Allocation (LDA) approach was used to analyze two corpora related to diabetes in order to identify key topics addressed in the texts[22]. LDA is a generative statistical model commonly used for topic identification in text data. Topics are defined as groups of representative words that aid in identifying the subject matter of the text. LDA represents documents as a mixture of latent topics, where each topic is characterized by a distribution over words[23]. Topic identification is crucial for clustering documents, extracting information from unstructured text, and selecting features. In this study, we aimed to uncover different topics, each represented by a combination of keywords, in the diabetes corpora to gain insights into the main themes discussed in the texts.

3.3. Sentiment Analysis

Finally, we did sentiment analysis of the text collected from online media sources related to diabetes. We classified the sentiments into three possible forms: neutral, positive and negative. This was realized by setting up a classifier on textual data with three classes.

Due to the lack of a labelled dataset for diabetes, another alternative was used to train the NLP models for sentiment analysis. Three different and famously used datasets are retrieved from the internet, namely: SST5 [24], Amazon Review [25], and Yelp Review [26]. SST5 is the Stanford Sentiment Treebank dataset consisting of 5 classes of sentiment on movie reviews; it is well-regarded as a crucial dataset and used as a primary benchmark dataset because of its capability to test an NLP model on sentiment analysis [27].To build the training and test data for our classifier, we used the publicly available datasets most frequently used by the NLP research community: SST5, Amazon Review and Yelp Review [24,25,26] A study of the latest literature in the field of textual data classification allowed us to conclude that pre-trained architectures of the Transformer type, such as BERT and some of its variants, achieve the best scores in terms of accuracy compared to classical architectures such as RNN or CNN[28].

Nine classification models were implemented by performing fine-tuning on pretrained models based on the BERT Transformer[29] .In this study, all models were implemented using the transformer library in Python, with the appropriate tokenizer selected depending on the specific model. Both ELMo and ULMFIT employ long short-term memory (LSTM) networks. In addition to the differing operational mechanisms of these two approaches, it is worth noting that the use of transformers enables parallelization of training, which is a crucial consideration when working with large datasets. In contrast to other models, BioBERT, Bio-ClinicalBERT, and PubMedBERT have been pre-trained on medical and clinical notes, as opposed to more generic corpora such as Wikipedia and English dictionaries. Devlin et al.(2019) proposes the following range of values are recommended: Batch size: 16, 32; Learning rate (Adam): 5e-5, 3e-5, 2e-5; Number of epochs: 2, 3, 4[29].

4. RESULTS

4.1. Dataset

The size of the corpus needed for a given task is determined by factors such as the intended use, computer processing speed, storage capacity, and the frequency and distribution of the linguistic features of diabetes in the corpus. The corpus for Indian Ocean and France contains 30,646 and 25,166 unique words , respectively, which is considered a good representation of the dataset. The table 1 gives an insight on the content of the corpora in the two different regions.

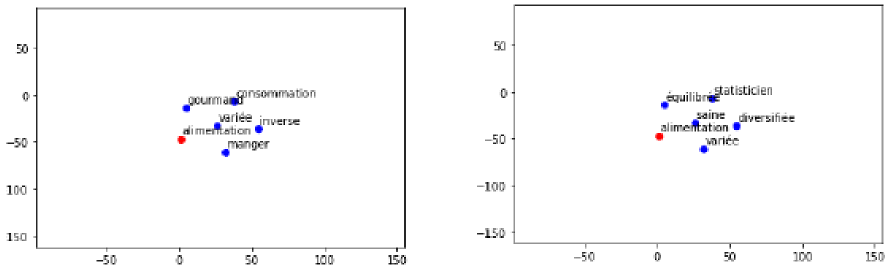
Table 1: Summary of Dataset

Variables	Indian Ocean Region	France
Number of Articles	613	607
Avg. Word/Sentence	27.8	21.8
Readability Index	14.18	12.94
Total number of words	493,744	25,166
Unique words	30,646	25,166

The vocabulary density and Readability index for Indian Ocean and France are 0.062 and 0.063 and 14.18 and 12.94, respectively. A low vocabulary density indicates complex text with many unique words, while a high ratio indicates simpler text with frequently reused words. A low density in this case also indicates that the corpus is well balanced. A readability index is an estimation of how difficult a text is to read, based on factors such as word lengths, sentence lengths, and syllable counts. The scores obtained indicate that the text is a bit difficult to read and is best suited for college graduates, with a score of 0-10 considered for professionals.

4.2. Word Association

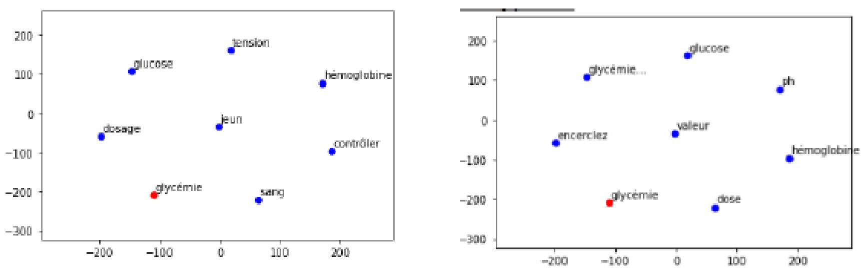
The results of applying Word2Vec to the corpus are illustrated in figures 2, 3, 4, 5. While some words association are common to both region, for alimentation (eaiting in english) in the indian ocean region the words ”diversifi´e”,”´equilibr´e” which means diversified and balanced respectively does not appear. For the next word ”glycemie”, which basically mean blood sugar for france the word ”tension”, meaning blood pressure, is does not appear in proximity. The information shown can be further interpreted by health care professionals and communication experts. The position of surrounding words in relation to the target word is crucial as it illustrates the semantic similarities between words, with closer words being more similar. The implemented code utilizes a lexicon of words related to diabetes to investigate how the disease is perceived and characterized. The number of similar words can be adjusted for any lexicon word used in the analysis.



(a) Indian Ocean

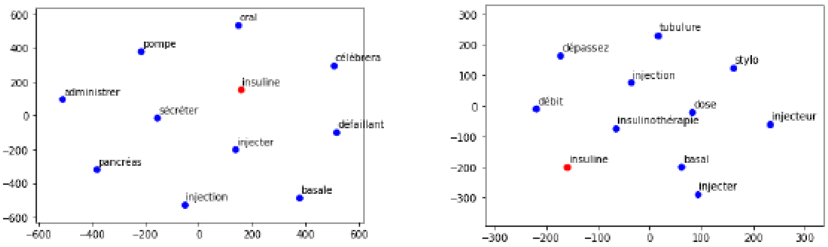
(b) France

Fig. 2: similar words to the word ‘alimentation’



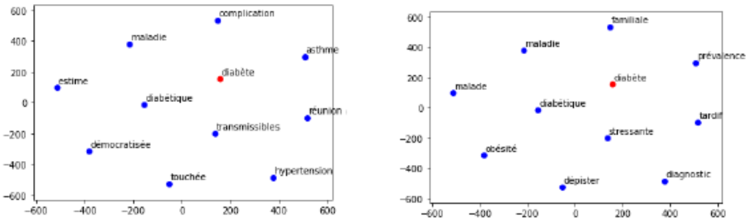
(a) Indian Ocean (b) France

Fig. 3: similar words to the words ‘Glyc´emie’



(a) Indian Ocean (b) France

Fig. 4: similar words to the words ‘Insuline’



(a) Indian Ocean (b) France

Fig. 5: similar words to the words ‘Diab`ete’

4.3. Word Analogy

Table 2 shows the results for word analogies for the two corpus. This give insight to information which would not have been possible without the application of NLP. For example we can see that in indian ocean (IO)

$$\mathit{regionsport} + \text{”}\mathit{alimentation}\text{”}(\mathit{eating}) - \mathit{hypertention}$$

is equal to ”manger” (eating) which in France it is equal to ”endurance” which is basically more oriented to resistance to effort in sport. In IO region the

$$\mathit{complication} + \mathit{malade}(\mathit{sick}) - \text{”}\mathit{immunitaire}\text{”}(\mathit{imune})$$

equals to amputation which is a major concern in Mauritius however in France it points to ”neophropatie” which is linked to renal diseases. These insights can again be better interpreted by healthcare professionals. Thhis technique is a power tool in medical field.

Table 2: Word Analogy example

Analogy	Indian Ocean	France
sport + alimentation – hypertension =	manger	endurance
diab`ete + r´enale – fruit =	c´ecit´e	insuffisance
complication + maladie – immunitaire =	amputation	n´ephropathie

4.4. Topic Identification

Latent Dirichlet Allocation (LDA) model was implemented to identify topics related to diabetes. The model was built with 20 different topics, where each topic is a combination of keywords, and each keyword contributes a certain weight to the topic. Table 3 provides a summary of the main subjects of the different topics around diabetes.

To evaluate the performance of the model, perplexity was calculated. Perplexity is a statistical measure that is used to determine the quality of a given topic model. A lower perplexity value indicates that the model is better at predicting the probability of the word in a given topic, hence a better model.

Table 3: Results of Topic Identification.

Topics	Indian Ocean	France
1	d´epistage	glycemie-insuline
2	ob´esit´e	diabete type
3	alimentation	sympto^me
4	consomation d’alcool	malade
5	patient	insuline
6	enfant	hypoglycemie
7	fruits	sucre
8	covid	alcool
9	vaccin	vitamin
10	ho^pital	Glucozor
11	´etude	enfant
12	sports	sommeil
13	madagascar	sports
14	alimentation	potassium
15	r´egime	grossesse
16	travail	chocolat
17	pharmacie	fruit
18	sucre	glycemie
19	traitement	alimentation
20	trouble	m´edecin

The analysis of the data from both regions revealed a high degree of similarity in the topics identified, with many common themes such as ”enfant”, ”alimentation”, ”alcool”, and ”fruit” being present in both regions. However, there were also some notable differences, such as the presence of topics related to ”grossesse”and ”sommeil” in France, but not in the Indian Ocean region. These disparities represent important areas for further investigation.

In order to evaluate the performance of the Latent Dirichlet Allocation (LDA) model used in this study, model perplexity was calculated. Perplexity is a widely used measure in text analysis that provides an objective means of determining the quality of a given topic model. The perplexity values obtained for Indian Ocean region and France regions were -9.494 and -9.496, respectively. It is worth noting that LDA is a popular tool for text analysis, providing both a predictive and latent topic representation of the corpus, however, it is equally important to identify if a trained model is objectively good or bad. The calculation of perplexity helps in determining the quality of the model. Overall, the results of this study indicate that LDA is a suitable tool for identifying similarities and differences in the topics of the data from these two regions.

4.5. Sentiment Analysis

For sentiment analysis a total of 30,000 training data and 6,000 testing data from the datasets SST5, Amazon Review, and Yelp Review were used. While there are numerous models available for this type of natural language processing (NLP) task, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), literature suggests that models based on transformers, such as BERT, and transfer learning have been observed to exhibit superior performance in comparison to these conventional models.

Table 4: Parameters used for training sentiment analysis

Models	Learning rate	Number of epochs	Batch size
BERT Based	2e-5	30	4
RoBERTa-Base	2e-5	30	4
ELMo	3e-5	10	16
ULMFIT	2e-5	10	6
OpenAI GPT2	2e-5	15	6
DistilBERT	5e-5	20	4
BioBERT	5e-5	4	16
Bio-ClinicalBERT	5e-5	4	16
PubMed BERT	5e-5	4	16

The learning rate hyperparameter controls the rate or speed at which the model learns. Usually, a large learning rate enables the model to learn faster, whereas a smaller rate will take significantly longer to train but may allow the model to learn a more optimal or even globally optimal set of weights. The number of epochs is a hyperparameter that outlines the number of times that the learning algorithm will work through the entire training dataset.

Due to limitations in computational resources available on Google Colab, which was utilized as the training platform for sentiment analysis, some models that underwent a higher number of epochs with a reduced batch size exhibited comparable results to models that underwent fewer epochs but were trained using larger batch sizes.Table 4 show the parameters used to train the various models.

Table 5 presents the results of training models using a combination of data from SST5, Amazon Reviews, and Yelp Reviews. As the focus of the study is diabetes data, incorporating these three datasets in the training process offers an advantage for the final testing phase, utilizing transfer learning on the custom diabetes corpus.

Table 5: Results of Sentiment Analysis

Models	Accuracy	Loss	Training time
BERT Base	91.6%	0.34	3 hrs 15 mins
RoBERTa Base	**92%**	**0.27**	**03 hrs 34 mins**
DistilBERT Base	90.2%	0.43	3 hrs 42 mins
Elmo	47% early stopping	1.19	1 hr 06 mins
ULMFit	89.2%	0.89	1 hr 45 mins
OpenAI GPT2	87.4%	1.02	1 hr 23 mins
Bio-ClinicalBERT	85.9%	0.62	2 hrs 09 mins
BioMed-PubMedBER	T 88.5%	0.68	2 hrs 42 mins
BioBERT	88.6%	0.65	2 hrs 39 mins

From the analysis of nine sentiment classification models, it was observed that RoBERTa outperformed the other models with an accuracy of 92%. The model underwent training for a duration of 3 hours and 34 minutes, resulting in a loss value of 0.27. The loss metric reflects the performance of the model after each iteration of optimization. A loss value less than 0.05 would indicate underfitting, while a value greater than 0.05 would indicate overfitting. Thus, a loss value of 0.27 is considered an acceptable outcome.

Consequently the diabetes corpus was trained using RoBERTa and the results are presented in Table 6 below:

Table 6: Results of Sentiment Analysis on Diabetes Corpus

Classes	Indian Ocean	France
Negative	294 (48%)	195 (32%)
Neutral	209 (34%)	254 (42%)
Positive	110 (18%)	158 (26%)

Table 6 reveals that a greater proportion of articles about diabetes in the Indian Ocean region are negative, at 48%, compared to those in France, which are 32% negative. This suggests that the Indian Ocean region has a higher frequency of discussion about the negative impacts of diabetes. In comparison, France has a higher proportion of neutral articles (42%), compared to positive (26%), and negative (32%), indicating that discussions in France are primarily focused on providing basic diabetes information.

5. DISCUSSIONS

Health literacy refers to a set of skills necessary for effective functioning within healthcare settings, while literacy skills are becoming increasingly important for functioning within society. Low literacy has been shown to have negative effects on both health and healthcare outcomes [30]. Designing precise communication materials is one factor that contributes to improving health literacy. However, to the best of our knowledge, no such survey has been conducted in the literature for the regions of Mauritius and France.

According to the International Diabetes Federation, the prevalence of Type 2 Diabetes is 22.6% in Mauritius and 5.3% in France [4]. This disparity suggests the need for collaboration in sharing knowledge between these regions. For instance, word association surveys on diabetes reveal that the word ”family” appears in the responses from France, but is completely absent in the responses from the Indian Ocean (IO) region. Similarly, the words ”obesity” and ”stress” are also absent in the IO region. Obesity is a significant risk factor for pre-diabetes and diabetes [31], and is a key target for the prevention and treatment of diabetes [32]. These findings highlight important factors that communication efforts should focus on, including stress management and obesity prevention.

Concerning sentiment analysis, a discrepancy is evident when comparing the negative, neutral, and positive sentiments regarding diabetes. Notably, the IO region emphasizes the presentation of diabetes’ negative aspects. Analyzing patients’ sentiments can facilitate rapid problem-solving and assist decision-makers in formulating effective plans for change [33]. This analysis emphasizes the critical need for health literacy experts and mass communication specialists in the IO region to reevaluate their strategies. Given the current high prevalence of diabetes in Mauritius and projected future increases, a multidisciplinary approach is required to effectively combat the disease and its complications, which pose a significant socioeconomic burden on developing countries.

6. CONCLUSION AND FUTURE WORKS

In conclusion, our study presents a pioneering approach utilizing Natural Language Processing (NLP) to analyze online media articles on diabetes, with implications for machine learning and healthcare research. The is a need to improve the data acquisition to include information which are shared by associations working on prevention and care of diabetes, including official communication done by public institution. On the use of NLP techniques we have demonstrated that the techniques is mastered and can be optimally applied. Our findings not only contribute to the existing body of work on NLP in the health sector but also demonstrate the potential of extending these techniques to electronic medical records and other non-medical domains. Moreover, we emphasize that combating the diabetes pandemic demands a multidisciplinary approach, encompassing prevention, education, communication, care, and management, involving not only healthcare professionals but also other stakeholders. Our study sheds light on the critical need for effective communication, education, and prevention messages delivered through online news articles, which remain a primary information source for many individuals in local and regional contexts. This research shows the importance of further efforts to ensure accurate and reliable information dissemination in the fight against diabetes.

REFERENCES

[1] L. J. F. Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser,and B. W. Hesse, “Online health information seeking among us adults: Measuring progress toward a healthy people 2020 objective,” *Public Health Reports*, vol. 134, no. 6, pp. 617–625, 2019.

[2] A. Boissonnet, M. Saeidi, V. Plachouras, and A. Vlachos, “Explainable assessmentof healthcare articles with qa,” in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 1–9.

[3] T. J. Leeper and R. Slothuus, “151How the News Media Persuades: Framing Effectsand Beyond,” in *The Oxford Handbook of Electoral Persuasion*. Oxford University Press, 06 2020.

[4] I. D. F. D. A. 10th edition scientific committee. (2021) Idf diabetes atlas 10thedition.

[5] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan,J. Marsh, J. Devylder, M. Walter, S. Berrouiguet *et al.*, “Machine learning and natural language processing in mental health: systematic review,” *Journal of Medical Internet Research*, vol. 23, no. 5, p. e15708, 2021.

[6] E. Kang, N. Song, and H. Ju, “Contents and sentiment analysis of newspaper articles and comments on telemedicine in korea: Before and after of covid-19 outbreak,” *Health Informatics Journal*, vol. 28, no. 1, p. 14604582221075549, 2022.

[7] S. Wang, F. Song, Q. Qiao, Y. Liu, J. Chen, and J. Ma, “A comparativestudy of natural language processing algorithms based on cities changing diabetes vulnerability data,” *Healthcare*, vol. 10, no. 6, p. 1119, Jun 2022. [Online]. Available: http://dx.doi.org/10.3390/healthcare10061119

[8] O. Oyebode and R. Orji, “Detecting factors responsible for diabetes prevalencein nigeria using social media and machine learning,” in *2019 15th International Conference on Network and Service Management (CNSM).* IEEE, 2019, pp. 1–4.

[9] ——, “Mediner: Understanding diabetes management strategies based on social media discourse,” in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom).* IEEE, 2021, pp. 1546–1553.

[10] P. SV, J. M. Lorenz, R. Ittamalla, K. Dhama, C. Chakraborty, D. V. S.Kumar, and T. Mohan, “Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people’s perspectives regarding covid-19 booster vaccine shots in india: Crucial to expanding vaccination coverage,” *Vaccines*, vol. 10, no. 11, p. 1929, Nov 2022. [Online]. Available: http://dx.doi.org/10.3390/vaccines10111929

[11] M. Canaparo, E. Ronchieri, and L. Scarso, “A natural language processing approach for analyzing covid-19 vaccination response in multi-language and geolocalized tweets,” *Healthcare Analytics*, p. 100172, 2023.

[12] M. S. Zulfiker, N. Kabir, A. A. Biswas, S. Zulfiker, and M. S. Uddin, “Analyzing the public sentiment on covid-19 vaccination in social media: Bangladesh context,” *Array*, vol. 15, p. 100204, 2022.

[13] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, “Diabetes on twitter: a sentiment analysis,” *Journal of diabetes science and technology*, vol. 13, no. 3, pp. 439–444, 2019.

[14] M. d. P. Salas-Za´rate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A.Rodriguez-Garcia, and R. Valencia-Garcia, “Sentiment analysis on tweets about diabetes: an aspect-level approach,” *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[15] I. De la Torre-D´ıez, F. J. D´ıaz-Pernas, and M. Anto´n-Rodr´ıguez, “A content analysis of chronic diseases social groups on facebook and twitter,” *Telemedicine and e-Health*, vol. 18, no. 6, pp. 404–408, 2012.

[16] Y. Liu, R. Stouffs, and Y. L. Theng, “Sentiment analysis on social media for identifying public awareness of type 2 diabetes,” in *The 54th International Conference of the Architectural Science Association (ANZAScA)*, 2020.

[17] K. Foley, D. McNaughton, and P. Ward, “Monitoring the ‘diabetes epidemic’: A framing analysis of united kingdom print news 1993-2013,” *PloS one*, vol. 15, no. 1, p. e0225794, 2020.

[18] N. F. Syafhan, G. Chen, C. Parsons, and J. C. McElnay, “Potential of uk and us newspapers for shaping patients’ knowledge and perceptions about antidiabetic medicines: a content analysis,” *Journal of Pharmaceutical Policy and Practice*, vol. 15, no. 1, pp. 1–11, 2022.

[19] ParseHub. (2022) The most powerful web scraper. [Online]. Available: https://www.parsehub.com/

[20] Octoparse. (2019) Easy web scraping for anyone. [Online]. Available: https: //www.octoparse.com/

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[23] ——, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] DeepAI. (2015) Stanford sentiment treebank dataset. [Online]. Available: https://deepai.org/dataset/stanford-sentiment-treebank

[25] M. Julian. (2018) Amazon review data. [Online]. Available: https://jmcauley. ucsd.edu/data/amazon/

[26] Y. Inc. (2019) Yelp open dataset. [Online]. Available: https://www.yelp.com/ dataset

[27] J. Wei. (2020) The stanford sentiment treebank (sst): Studying sentiment analysis using nlp. [Online]. Available: https://towardsdatascience.com/ the-stanford-sentiment-treebank-sst-studying-sentiment-analysis-using-nlp-e1a4cad03065

[28] S. Gonza´lez-Carvajal and E. C. Garrido-Mercha´n, “Comparing bert against traditional machine learning text classification,” *arXiv preprint arXiv:2005.13012*, 2020.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training ofdeep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[30] N. D. Berkman, T. C. Davis, and L. McCormack, “Health literacy: what is it?”*Journal of health communication*, vol. 15, no. S2, pp. 9–19, 2010.

[31] A. Boles, R. Kandimalla, and P. H. Reddy, “Dynamics of diabetes and obesity: Epidemiological perspective,” *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1863, no. 5, pp. 1026–1036, 2017.

[32] M. S. Rendell, “Obesity and diabetes: the final frontier,” *Expert Review of Endocrinology & Metabolism*, no. just-accepted, 2023.

[33] L. Abualigah, H. Alfar, M. Shehab, and A. Abu Hussein, *Sentiment Analysis in Healthcare: A Brief Review*, 01 2020, pp. 129–141.

AUTHORS

Shaad Toofanee is a Senior Lecturer at the Universit´e des Mascareignes in Mauritius(UDM). He is presently pursuing a Phd in the field of Artificial Intelligence at the Universit´e de Limoges France (UNILIM)in the research lab XLIM, UMR CNRS 7252. His research interests are using Mahcine Learning in the field of health and more precisely diabetes prevention, management and education. He is presently investigating the use of vision transformer for image and Transformers for Natural Language Processing. He is co-director of the programme Master in artificial Intelligence and Robotics.

Sabeena Dowlut is a Senior Lecturer at the Universit´e des Mascareignes in Mauritius(UDM) and Head of depart of Applied Computer Science. She has a Phd in Health Litteracy from Universit´e de la R´eunion. She is also co-director of a Master programme in Health and AI which will be starting in September 2023.She is also member of R´eseau francophone de litt´eratie en sant´e(REFLIS) and a member of the Eco-´ nomic and Scientific Committee of AUF for the Indian Ocean and African Austral region. She is presently co-supervising a Phd thesis in the field of IoT and Health.

Nabeelah Pooloo graduated with a Bsc. Honours degree in software engineering degree from Universit´e des Mascareignes. She was successfully selected for a fully funded government of mauritius scholarship for a Master degree in Artificial Intelligence and Robotics which was jointly offered by Universit´e des Mascareignes and University of Limoges. She successfully completed her research internship at XLIM laboratory, UMR CNRS 7252 (University of Limoges).

Karim Tamine is an Associate Professor / Researcher at thein the research lab XLIM, UMR CNRS 7252 (University of Limoges). His research work focuses on the use of Artificial Intelligence methods in various fields such as computer graphics, security and quality of service in dynamic communication networks. He is the main resource person in artificial intelligence for the setting up of a master degree at Universit´e des Mascareignes Mauritius. He has supervised 8 phd thesis and he is presently supervising 2 phd students.He has taken an interest in the field of application of Machine Learning in healthcare.

Damien Sauveron is Professor/ Researcher at the XLIM laboratory (University of Limoges). He is also presently dean of the faculty of science and technology.His research interests are related to Smart Card applications and security (at hardware and software level), RFID/NFC applications and security, Mobile networks (e.g UAV fleets) applications and security, Sensors network applications and security, Smart home applications and security, Internet of Things (IoT) security, Cyber-Physical Systems security, security of Distributed Objects and Systems and security evaluation/certification processes.