1-1-2020

Tweets classification and sentiment analysis for personalized tweets recommendation

Asad Masood Khattak
Zayed University, asad.khattak@zu.ac.ae

Rabia Batool
Zayed University

Fahad Ahmed Satti
Kyung Hee University

Jamil Hussain
Kyung Hee University

Wajahat Ali Khan
University of Derby

See next page for additional authors

Hindawi
Complexity
Volume 2020, Article ID 8892552, 11 pages
https://doi.org/10.1155/2020/8892552

Research Article

Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation

Asad Masood Khattak,[1] **Rabia Batool**,[1] **Fahad Ahmed Satti**,[2] **Jamil Hussain**,[2] **Wajahat Ali Khan**,[3] **Adil Mehmood Khan**,[4] **and Bashir Hayat**[5]

[1]*College of Technological Innovation, Zayed University, Dubai, UAE*
[2]*Department of Computer Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea*
[3]*College of Engineering and Technology, University of Derby, Markeaton Street, Derby DE223AW, UK*
[4]*Institute of Information Systems, Innopolis University, Innopolis, Russia*
[5]*Institute of Management Sciences, Peshawar, Pakistan*

Correspondence should be addressed to Asad Masood Khattak; asad.khattak@zu.ac.ae

Received 12 August 2020; Revised 23 November 2020; Accepted 1 December 2020; Published 17 December 2020

Academic Editor: Atif Khan

Mining social network data and developing user profile from unstructured and informal data are a challenging task. The proposed research builds user profile using Twitter data which is later helpful to provide the user with personalized recommendations. Publicly available tweets are fetched and classified and sentiments expressed in tweets are extracted and normalized. This research uses domain-specific seed list to classify tweets. Semantic and syntactic analysis on tweets is performed to minimize information loss during the process of tweets classification. After precise classification and sentiment analysis, the system builds user interest-based profile by analyzing user's post on Twitter to know about user interests. The proposed system was tested on a dataset of almost 1 million tweets and was able to classify up to 96% tweets accurately.

1. Introduction

In the last decade, social networks have witnessed multifold advancements due to the rapid digitization of the service industry and other advancements in the field of information technology. A plethora of information sharing platforms and the increased connectivity with the Internet [1] have also led to a change in the general perspective of networking, socialization, and personalization [2]. For the month of December 2018, an average of 1.52 billion users were active on Facebook daily [3]. This is besides auxiliary services offered by Facebook, such as WhatsApp, Messenger, and Instagram, each of which has over 1 billion active users, per month [4]. Similarly, as identified from third-party reports, other platforms, such as YouTube owned by Google, iMessage by Apple, and WeChat by Tencent, are also a part of the, no longer elite, 1 billion-per-month-active-user-club. More significantly, three out of every four adult Internet users are now actively utilizing at least one social network platform [5]. From a pure technological point of view, this enhanced connectivity has created unique challenges and opportunities [6, 7] by allowing the users to not only consume services but also to share their experiences, feelings, and thoughts. One of the most impactful and emerging social networks is Twitter, which allows its users to broadcast the latest (personal, communal, national, or international) events in the form of short messages, "tweets," which are typically comprised of text, audiovisual content, and/or links to external websites [8, 9]. Twitter is playing a key role in many fields such as social marketing [10], election campaigns [11], academia [12], and news. Hashtags (words identified by the symbol #) form a key part of any tweet, allowing public content to be categorized and made searchable for users. This allows the hashtag(s) to enrich the

shared content and enable valuable analysis leading to discovering new insights and trends. In terms of information discovery and knowledge creation, this plethora of user created content allows the application of sentiment analysis, which aims to provide an automated mechanism for determining the writer's attitude towards the subject or its overall contextual polarity [13]. These insights are especially useful for digital marketing, allowing organizations and in some cases governments (such as during the Arab Spring [14]) to monitor and measure social media and gain actionable business/social intelligence, allowing to understand how people view their brands, products, and services and to improve brand visibility.

In the same manner, social media is now playing an active role in improving healthcare service delivery [15]. Shifting to a more user-centric approach, social media enables near real-time information flow, which in turn enables immediate interventions for individuals and communities in hospitals, at clinics, or at homes [16]. For example, in a survey [17], the authors reported that search for health information stood out as the third most popular online activity. Today, patients, irrespective of their age, gender, or socio-economic standings, are relying on the web to find healthcare information related to their particular needs [18, 19]. Additionally, patients can now make more informed decisions by examining the experiences of their peers in terms of symptoms, reactions, and treatments related to a particular disease, thereby bridging the communication gap between the patients and healthcare providers [20]. In addition, healthcare organizations can also take benefits by finding the timely response of problems and monitoring the user's behaviors, conditions, and feelings in between their visits [2]. Keckley and Hoffmann [21] studied online social networks to analyze their effect on patient health and found that people get more benefit while sharing their data on social networks such as PatientsLikeMe portal [22]. This virtual connectivity can provide many benefits, such as improving medication adherence, pharmacovigilance [23], reduction in side effects, enhanced community support, improved epidemiological analysis [24], and generally better healthcare services. Consequently, it is safe to say that healthcare benefits are directly related to social reachability [25]. According to PwC Health Research Institute [26], almost 90% users in the age of 18–24 were willing to share their health information on social networks.

However, such large use of social media has also introduced the problem of information overload. With an overwhelming amount of data on social media, users find it difficult to get personalized and concise information. Short and noisy text on social media also makes it hard to understand full context and classify data. In this paper, we propose a framework for providing personalized recommendations to the user by analyzing his health interest on social networks. While this work can be generalized in many domains, the research work presented henceforth is focused on processing healthcare data and information.

The proposed classification and sentiment analysis system uses a semantic structure, important keywords, and opinion words from tweets to monitor user interests and then generates personalized healthcare and wellness-related tweet recommendations. These personalized tweets consist of publicly available content which is precisely preclassified by our system. For tweet classification, the proposed system uses a domain-specific seed list which helps to decide which category a particular tweet belongs to. After classification, the proposed system also applies a lexicon-based sentiment analysis approach to extract topic level sentiments in tweets. To increase the accuracy of tweet analysis, the proposed system also uses synonyms with keywords. The proposed model performs more precise analyses of tweets enriching temporal patterns and semantics of keywords which optimize filtering result and help to extract more knowledge from tweets. For testing of profile generation, we collected 6000 tweets of users and generated user profile by extracting health-related keywords, entities, and sentiments. For classification, the system was tested on almost 1,000,000 tweets of different categories. Due to our preclassification strategy and other significant improvements, our current model showed an accuracy of 96% for tweet classification, which is significantly better than our previously published approach, with an accuracy of 89.5% [27]. The proposed system also measured how much information for one category can be extracted from other categories which were ignored by keyword-based search from tweets.

The main contribution of the presented work is complete design and implementation of a personalized recommender system for a user based on his temporal social media history. The proposed system does not just rely on keyword-based interest but it also takes user's temporal sentiments into account. The syntactic and semantic analysis of tweets leads to more complete profile generation and tweet classification.

The rest of this research paper is structured as follows. Section 2 discusses related work closely aligned with our work. In Section 3, we present the theoretical foundations of the proposed platform and its components, followed by Section 4, which briefly describes our implementation strategy and presents the evaluation results of the proposed system. In the end, Section 5 concludes the research work and highlights future work.

2. Related Work

Social media analytics is an active, interdisciplinary research field, which has enabled the researchers to gain unique perspectives into human and data behaviors. The volume and variety of this largely unstructured data, produced at high velocity, has led to the development of many tools and technologies for extracting or rather enhancing the value of social interactions. Yet, there still remain many challenges in terms of identifying relevant data, tracking actions and reactions, increasing the veracity of data, optimization of data storage, data processing and visualization of information, extracting hidden patterns, and closing the data to knowledge loop [28]. A key task for researchers pursuing applied research in this field is to not only identify the techniques used for converting data to information and subsequently knowledge but also to look at its impact [29]. Twitter, along with its streaming API, and a large open (in

terms of keeping their tweets public) user base has further enabled the monitoring and analysis of a rich gold mine of data produced via a novel information propagation strategy.

One of the more recent works in terms of analyzing tweet propagation, for prominent Mexican political figures, through the utilization of visual aids and pattern recognition approaches has been laid out by [30]. In this work, the authors collected tweets from six prominent Mexican politicians, their mentions, retweets, and favorites to their tweets. By applying sentiment analysis followed by a contrast pattern-based classifier working on 124 extracted (5 nominal and 119 numerical) features, the authors were to quantify impact of tweets based on their propagation patterns. In an earlier approach, as presented by [31], the authors utilized social features (such as number of followers, favorites, and others) and tweet features (such as number of hashtags, tweet length, and others) to predict the likelihood of a tweet being repropagated (also known as retweeting). In this work, the authors used a passive-aggressive algorithm for automated categorization of tweets. The performance of their model was slightly higher than manual categorization by human subjects. Tweet categorization is also important to identify relevant data for early responders, immediately after a disaster event. Li et al. [32] have built on earlier works and presented a supervised Naive Bayes model, along with an iterative self-training strategy which is able to provide good results. However, the presented results are from a controlled environment (CrisisLexT6-labelled data set, covering 6 disasters between Oct 2012 and July 2013), and its application in live environment would require a lot of data preprocessing.

A use case of such categorization is to build recommendation systems, which can provide a more personalized experience to the users. A basic URL recommendation system based on the user tweets, topic interest models, and social voting was introduced by Chen et al. [33]. Using 12 voting algorithms and feedback from 44 users, the authors were able to provide a basic platform for future recommendation systems based on Twitter data. Abel et al. [34–36] analyzed user modeling for presenting personalized news recommendation and improved the semantic of Twitter activities by enriching news items with tweets. The work used methods including topic-based, entity-based, and hashtags to analyze user modeling. They also focused on temporal pattern extraction in users' profile. Piao and Breslin [37] analyzed user modeling strategies by incorporating categories, classes, and connected entities from DBpedia for extending user interest profiles and found that their proposed method significantly outperforms existing approaches in the context of link recommendations. A dynamic user modeling-based recommendation system was proposed by Deng et al. [38] to integrate information extracted from tweets and the video ranking system employed by Youtube based on the same user's profile. This strategy greatly enhanced the relevancy of the video recommendations. Celik et al. [39] identified the semantic relationship between Twitter entities to provide mediation among the same, thereby allowing the users to access the relevant content of their interest. Balabanović and Shoham [40] proposed a system to build user profile by combining both collaborative and content-based recommendation techniques. In content-based recommendation systems, user preferences are considered for providing recommendations. On the other hand, in the collaborative recommendation, the system identifies users with similar taste to that of the given user and provides recommendation based on this similarity.

Another popular use case of data analytics on Twitter is sentiment analysis. Yi et al. [41] presented a model to extract only subject-based sentiments from tweets by extracting topics and sentiments, followed by an application of a mixture model to detect relations between them. Similarly, Nasukawa and Yi [42] identified sentiment related to the particular subject using natural language processing techniques. The novelty of their approach was based on Markov model-based tagger for recognizing part of speech, followed by statistics-based techniques to identify sentiments related to a subject. Godbole et al. [43] introduced a system to determine public sentiment, and its variation over time, for news and blog entities. Using synonyms and antonyms, the authors were able to find a path between positive and negative polarity and increase seed list.

Some of the other popular use cases include improved search, improved tweet contents, and predicting election outcomes. Reviewing studies catering to these use cases is an important tool for identifying the techniques, which can help improve the impact and effectiveness of the recommendation system. Guo and Lease [44] proposed a novel ranking model, for enriching the search functionality on Twitter, with personalization and content analysis. Clark and Araki [45] introduced a text normalization technique to categorize errors and informal language used on social media into different groups, followed by natural language processing techniques to correct common phonetic and slang mistakes. On the contrary, Laniado and Peter [46] applied hashtags on Twitter and demonstrated mappings of fifty percent hashtags to entities in freebase. The system was categorized into four dimensions: frequency, specificity, consistency, and stability to assess hashtags as strong identifiers. Lösch and Müller [47] proposed a method to associate hashtags with encyclopedia entities. Their system used Wikipedia entities as a description of hashtags in microblogging service to understand the actual context of hashtags. Tumasjan et al. [48] analyzed Twitter as a source of predicting elections. They used the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation. They used LIWC 2007 [49], a text analysis software, which uses a psychometrically validated dictionary for identifying and assessing the emotional, cognitive, and structural components of given text samples. The authors used 12 dimensions including past and future orientation, positive and negative emotion, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money to extract political sentiments from this data.

In this paper, we are providing the users with personalized health-related profiling and aggregated sentiment analysis using precisely classified data and sentiments. We

propose a novel approach for analyzing the behavior and lifestyle of individuals by monitoring patient's self-reported data and social posts. The archivist is a service that finds and archives tweets using Twitter search API. It helps the user to get real-time trend information on Twitter [50]. Our model uses the archivist to collect Twitter data and process them using natural language processing techniques to extract knowledge and sentiments from tweets. Twitter contains a lot of information; however, the proposed model focuses on how the information is filtered precisely to provide personalized knowledge to users.

3. The Proposed System Architecture

Twitter is a popular social media platform that enables users to post short texts, images, and videos of personal and/or collaborative nature. This data provides a unique insight into the user's personality. Of particular interest to our research work, are the user's interests and emotions, which are used by our proposed system to build a user profile and then provide personalized data/services to similar users. Our proposed system, as shown in Figure 1, consists of two modules and integrates with Twitter as a plug-in application. The first module builds user health profile by extracting the user's profile information, health interests, and emotions enriched with temporal patterns. To achieve the objectives, Alchemy API [51] is used for the extraction of user's interests from the free text (tweets). The API processes unstructured text using natural language processing techniques and machine learning algorithms to produce keywords, entities, concepts, and the sentiment of the user in relation to these (keywords and entities). The second module collects public data from Twitter and precisely classifies it to recommend users with personalized data based on their generated profile. To classify tweets and extract topic level sentiments, the system analyzes tweets using domain-specific seed words, opinion words, n-gram generator, POS tagger, synonym binder, and dependency parser. Seed words and opinion words are enriched by synonyms to increase accuracy of classification.

3.1. Data Manager. Data manager acts as a plugable interface to Twitter, which internally utilizes a data fetcher to acquire streaming data. These data are received in XML format, a sample of which is shown in Figure 2. Each tweet is encapsulated in a structured format, containing the username of the person tweeting, timestamp of the tweet, textual content of the tweet, IT unique identifier, any associated image, and other information. Using a DOM parser, we parse this XML corpus to extract the username, tweet date, status, tweet ID, and image fields. We then apply text preprocessing on the tweet text (status field) to convert the raw data into meaningful information. The main aim of this step is to convert abbreviations and slangs, contained in the tweets, into their formal counterparts. This aim is set to alleviate the tweet behaviorisms, which have informally encouraged the use of abbreviations (such as "*plz*" instead of "*please*" and "*gud*" instead of "*good*") and other slang words [52], by Twitter users to save time and space. Users can also repeat characters in words to emphasize a particular word (such as using "*Plzzz*, as shown in the second tweet in Table 1"). Such words represent noise in data, since it affects the knowledge extraction process.

The data preprocessor module achieves this aim by utilizing a repository of 1300 slang words to remove this noise. As a result of this process, the resulting data are free of most commonly used (on social media) slang and abbreviated words. Additionally, the spell checker module uses jazzy (Java-based spell checking API) to correct any spelling mistakes from the data. The final data produced by the data manager is very rich and can be used by the consuming services to build a user profile and extract knowledge.

3.2. Profile Builder. This submodule extracts useful information from tweets and maintains temporal history to build user health interest-based profile. Profile builder extracts the user's interests by using Alchemy API. It accepts unstructured text and obtains knowledge by exposing the semantic richness hidden in posts using named entity and sentiment related to those entities. System stores extracted keywords, entities, and user sentiments in the user's profile repository for future use. Table 1 shows a sample of the keywords, entities, and associated sentiments extracted by profile builder using the IBM Watson Natural Language Understanding module (Alchemy API). For instance, the tweet "I feel my high blood pressure is at an unsafe level every time I'm at work. It's seriously going to give me a depression one of these days" when processed through this API shows "high blood pressure" as the most relevant keyword with the highest confidence score of 0.99206. Similarly, the highest rated concept against this tweet is "hypertension" with a score of 0.915043. The overall sentiment associated with this keyword is negative with a confidence score of −0.96. Similarly, the other sample tweets with their corresponding keywords, entity concepts, and entity sentiments are shown in Table 1, along with their score in parentheses. For each of these attributes, we have selected the top one keyword, concept, and sentiments, with respect to their relevance in the text. It is also pertinent to note that not all entities are correctly identified, as in the case of the third example in the table "Wide awake, I've got a headache and work in the morning" which has the correctly identified keyword "headache" with a score of 0.71, but an unrelated concept "2006 singles" with a confidence score of 0.86%. We do not disregard this incorrect conceptualization, which only slightly affects that overall accuracy, as will be shown in the result section.

After extracting this information from tweets, profile builder searches for the temporal patterns of user interest, e.g., in the morning, the user is usually interested in the blood sugar level, while in the evening, the user usually talks about insulin and diet. If same pattern appears more than two times, profile builder attaches temporal information with the knowledge extracted to use it for data recommendations. All the extracted data and temporal information are then stored in the database.

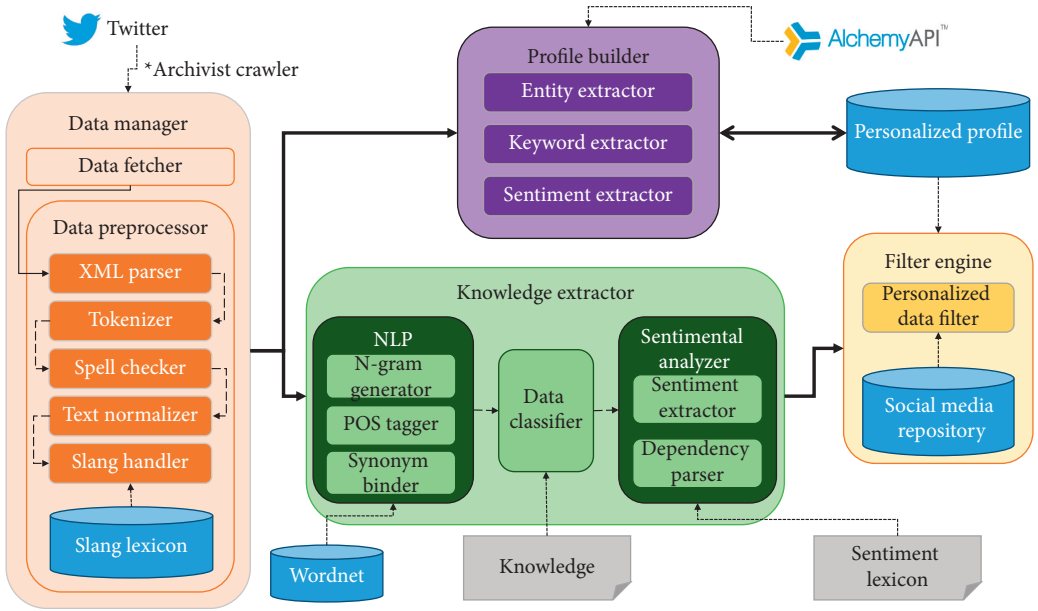

Figure 1: The proposed system architecture for profiling and tweet recommendation.

<Tweet
TweetStatus = "Unapproved," Username = "...", TweetDate = "2015-06-12T16:57:47+09:00"
Status = "I love new Diet PSPSI and hate aspartame."
TweetID = "…"
Image ="http://a0.twimg.com/profile_images/2569014433/IMG00354-20120526-1717_normal.jpg"
/>
<Tweet
TweetStatus = "Unapproved," Username = "…", TweetDate = "2015-04-12T16:59:47+09:00"
Status = "I am Diabetic. Here's how it works. My insulin pump and continuous glucose meter (CGM). Plzzz help me"
TweetID = "... ..."
Image = "http://a0.twimg.com/profile_images/2569014433/IMG00354-20120526-1717_normal.jpg"
/>

Figure 2: Sample Twitter data collected by the data manager.

Table 1: A sample of knowledge extracted by profile builder, with confidence scores shown in parenthesis.

Tweet	Keyword	Entity		
		Concept	Text	Sentiment
I feel my high blood pressure is at an unsafe level every time I'm at work. It's seriously going to give me a depression one of these days	High blood pressure (0.99204)	Hypertension (0.91) Health condition	High blood pressure Depression	Negative (−0.96) Negative
I am diabetic. Here's how it works. My insulin pump and continuous glucose meter (CGM). Plzzz help me	Insulin pump (0.996333)	Insulin (0.96)	Diabetes	Negative (−0.54)
Wide awake, I've got a headache and work in the morning	Headache (0.71)	2006 singles (0.858618)	Headache	Negative (−0.8)
I am healthy and feeling good after having high blood pressure now	High blood pressure (0.981841)	Hypertension (0.915043)	High blood pressure	Positive (0.83)

3.3. Knowledge Extractor. Knowledge extractor module consumes the processed tweets, coming from the data manager in order to apply natural language processing and sentiment analysis techniques to precisely classify them. In particular, the proposed system uses the Stanford Part-of-Speech (POS) tagger, dependency parser, four-gram, and a synonym binder to classify the tweets. The tags identified by the Stanford POS tagger are used to extract synonyms from

WordNet. Additionally, the synonym binder helps improve the accuracy of classification by binding synonyms from the seed list with each noun word. This binder is based on the WordNet dictionary, which also allows us to identify the contextual meaning of the present words. Jaws API [53] provides the synonym binder with an external interface to WordNet. For example, the word workout is not present in our seed list; however, its bound synonym exercise does exist. The synonym binder also handles other problems related to word structure as well. For example, it can convert plurals to singulars, thereby binding calories with calorie and exercises with exercise. Sentiment analyzer uses sentiment lexicon to extract positive, negative, and neutral sentimental words from these enriched tweets. For positive and negative sentiments, the system uses the list of 6800 words from [54]. In addition, for neutral classes, a list of neutral keywords is built after analyzing tweets.

The proposed system classifies tweets based on the knowledge extracted from them. This classification process is dependent on the seed list, which is used to identify the particular category that a tweet belongs to. In this research work, we have focused on the healthcare domain by keeping the most frequently used healthcare and wellness terms in our seed list. The classified data are stored in a knowledge-base for improving accuracy and future use.

Once the proposed system has classified and detected sentimental words from tweets, the Stanford dependency parser was used to identify the relation between the extracted categories and sentimental words. This helps the system to find topic-based sentiments in tweets. The proposed system uses dot, exclamation mark, and hyphen as sentence boundaries for splitting tweets into sentences if there are multiple sentences in a tweet. Typed dependencies are grammatical relations between words which help to decide either a sentiment belongs to a specific word or not. It also helps for extracting multiple sentiments from a tweet. Figure 3 shows how dependencies are used to find topic-based sentiments. Dependency parser also helped to find negation of any sentimental words to inverse its value, e.g., in tweet "I don't like the taste of that medicine" has the negation of a positive word "like." Without considering negation, the system was not able to link negative sentiment to "taste."

3.4. Filter Engine. Filter engine processes classified tweets using personalized profile and aggregate sentimental result to recommend the user with relevant data. While generating data recommendation, filter engine also incorporates temporal patterns extracted by profile builder to generate more valuable, time-specific recommendations. Figure 4 shows the positive, negative, and neutral sentiments associated with the various common drugs used by diabetic patients and mentioned in their tweets. This sort of filtering can enable the physicians and caregivers to optimize drug delivery by incorporating the patient sentiments in their medicine prescription process. This could enable a positive impact on the medication adherence by the diabetic patient. Figure 5 shows another use case of the filter engine's application, whereby the diabetic patient is shown relevant tweets based on similar keywords and sentiments to reenforce constructive dialog and create a virtual support system for the diabetic patient. Through this approach, the patients can obtain useful information related to their disease and others' experiences on different kind of insulin, drugs, or medical tests.

4. Implementation and Result

While the presented approach can be generalized to any domain, in this research work, we have extended our previous approach, presented in [27], to extract healthcare knowledge from publicly available tweets, providing recommendations for diabetes. In order to realize the proposed framework, we have used Java and other open APIs to create an application which amalgamates the data curation service, knowledge extraction service, user profile building service, and filter engine into the proposed recommendation system. These services are briefly explained in the following subsections.

By applying seed list-based classification and sentiment analysis, the system was able to recommend personalized diabetes-related tweets to users. The seed list was generated using the work presented in [55, 56]. In order to overcome redundancy problems and formatting issues, Google Refine is used. To calculate the accuracy of our proposed system, we have used seed list for diabetes for tweet filtration. By integrating our proposed system with Twitter, the user would be able to get precisely classified and personalized data with sentiment value. Moreover, this tweet data is useful for clustering, trend analysis, and recommendations as well. The details of the data collection process, our experiments, and their results are as follows.

4.1. Data Collection. Archivist tool has been used to scrawl a specific set of tweets for all the keywords presented in Table 2. Table 2 also shows the number of extracted tweets, along with their classification accuracy when using only n-gram and when using n-gram with synonyms.

To generate user profile, we analyzed tweets of 100 users and collected 6000 tweets related to diabetes which helped to build user profile. Some collected tweets for profile generation could not provide any information about user health interests, so the system ignored them and used only those tweets which helped to generate user's health profile.

The seed list of diabetes-related terms has been generated by utilizing the work presented in [55, 56]. This list was then divided into two parts, by using natural language processing to classify diabetes-related terms, based on their definition in the original source. As a result, 417 terms have been classified into categories, such as test, condition, body cell, diabetic study, professional, devices, medicine, and others (not to be confused with the well-defined category "other"). For example, "hyperinsulinemia" was defined in the seed source, as "a condition in which the level of insulin in the blood is higher than normal caused by overproduction of insulin by the body." The proposed system classified it as a "condition"

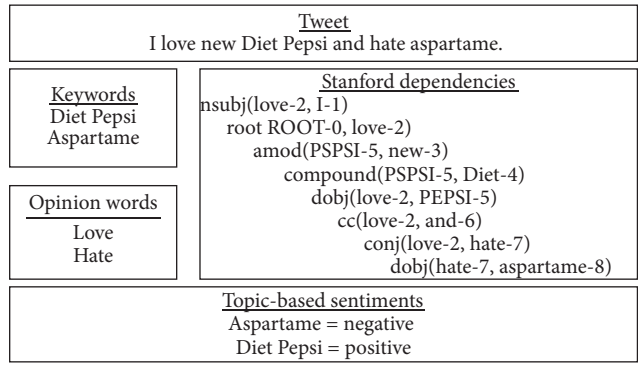

Figure 3: Dependencies from tweet for topic-based sentiment analysis.

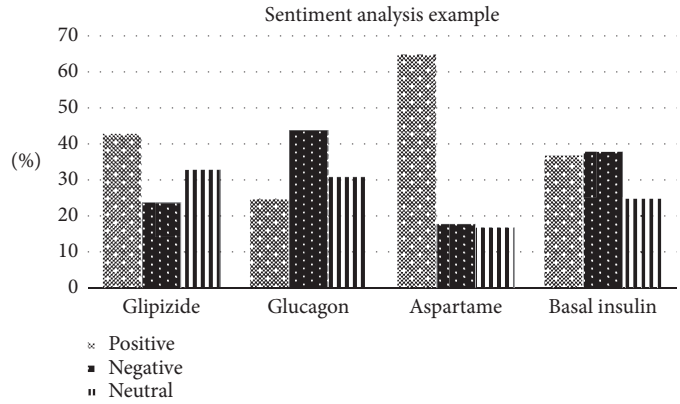

Figure 4: Sentiment analysis results for a diabetic person.

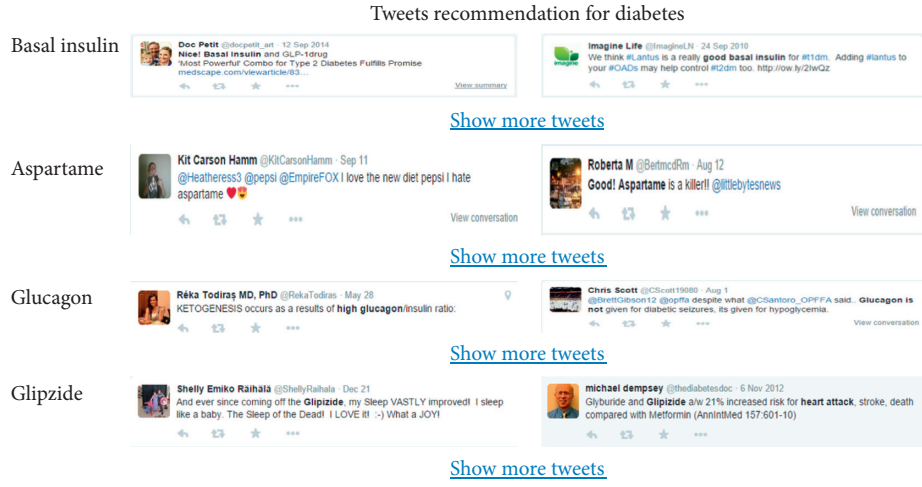

Figure 5: Data recommendation to diabetic persons.

term. Our system was able to classify 80.5% of the terms, leaving only 81 terms, which were labelled as belonging to the "other" category.

For sentiment analysis, the proposed system used the list of positive and negative sentiments which is composed of 6800 words from [54]. For neutral class, we manually build a list of 30 keywords.

4.2. Testing. Almost six thousand tweets were used to generate user health profile. By using Alchemy API, this system extracted all important keywords, entities, and sentiments from tweets. This information is used to build user profile which helped to provide the user with personalized data recommendation. The data recommendation is precisely classified data with public sentiment analysis.

Table 2: Diabetic tweets classification.

Category	Total	Classified as diabetic	
		4-gram (%)	4-gram + synonyms (%)
Diabetes	94992	95	96
Blood pressure	31659	95	95.60
Diet	37738	8.50	10.60
Medication	16997	4.30	5.60
Parkinson	6503	3.80	5.20
Food	42415	2.06	4.70
Education	245317	0.90	2.50
Dengue	5200	0.80	6.10
Pain	109067	0.50	1.90
Technology	110572	0.36	1.30
Entertainment	136308	0.20	1.05
Earth quake	103632	0.17	1.60
Movie	30943	0.10	1.20

Spell checker also improved system performance as social media data have spelling and typo errors.

The proposed system has processed almost one million tweets of different categories for testing and verification of classification and sentiment analysis. By considering only four-gram, from all categories, 129,839 diabetes-related tweets were successfully classified. However, when the proposed system was employed in full, which uses four-gram and the synonym binder, 142,285 diabetes-related tweets were classified, from all categories. This is because the synonym binder binds the context of words from tweets, which improves the categorization process. By applying preprocessing and then semantic and syntactic analysis, system accuracy has reached up to 96% for diabetes-related tweets, as shown in Table 2. The system used n-gram model with synonym binder to achieve this accuracy. Diabetes-related tweets from other categories decreased information loss and increased the quality of sentiment analysis. Simple keyword-based search from Twitter is not able to provide all the related information for a specific category. This can be greatly enhanced by using a seed list, which would enable the retrieval of information related to the keyword. In the legacy search case, the term "diabetes" would only return tweets, containing this keyword. However, using the seed list to perform an advanced search can also return additional information by retrieving those tweets, which do not explicitly contain this keyword but are still of interest to the diabetic patient or the caregiver, for example, "Morning walk is very helpful to maintain blood glucose." This tweet is not filtered when we search Twitter for diabetes; however, the proposed system has successfully classified this tweet as a diabetes-related tweet.

Dependency parser has helped the proposed system to find an accurate relationship between sentiments and classes. It has also helped to find multiple sentiments for multiple classes from a single tweet. Figure 3 shows how the proposed system has extracted topic-based multiple sentiments from a single tweet. At first, sentimental words and topics were extracted, but it was not clear which sentiment is related to which topic. So, the system used a dependency parser to bind sentiments with the topic. Dependency parser also helps the system in negation detection, e.g., "neg (good, not)" shows that "good" is negated. Negation inverts opinion of the sentimental word from positive to negative and vice versa. Figure 4 shows the sentiment analysis of the tweet data generated for a diabetic person. It shows that 37% tweets about basal insulin are positive, 38% are negative, and 25% have neutral sentiments. The figure shows that the majority of sentiment for glucagon is negative. These results help the user not only to find related tweets but also aggregated sentiments. Through the application of advanced natural language processing techniques, such as topic modeling, keyword extraction, and sentiment analysis, the classification accuracy is greatly improved. Figure 6 shows comparison of the proposed system with existing technique [27]. It shows 6.5% performance improvements, from existing technique, in terms of accurately classifying tweets related to diabetes and 22.8% improvement on classification for blood pressure.

Additionally, the proposed system addresses a key use case of information loss, caused by a legacy keyword-based search engine. Twitter search can be greatly enhanced by using seed lists and short text classification to extract a larger set of related information, without increasing the cognitive load on the user. Table 2 shows the effectiveness of using this process for extracting information related to diabetes. Information diffusion varies in each category; while 10.6% tweets from the diet category and 6.1% tweets from dengue contain valuable information about diabetes in the blood pressure category, we found 95% of tweets containing content related to diabetes. Legacy keyword search on Twitter was not able to extract these tweets. It is also important to note that the information collected through this process is not unique, and as we found out, there is an

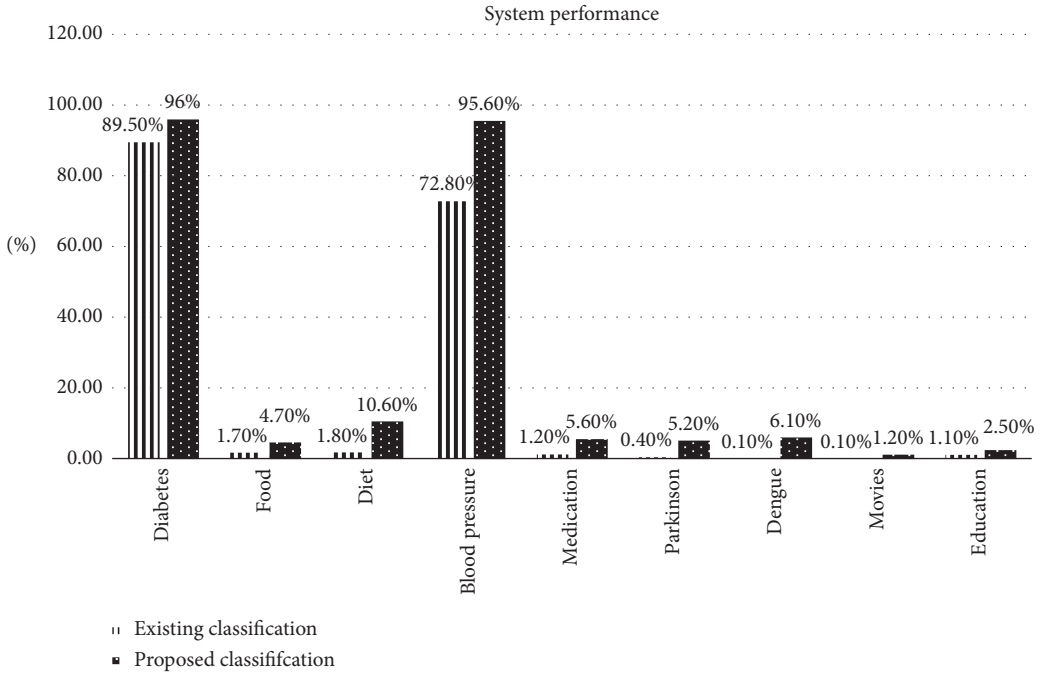

Figure 6: Performance comparison with existing technique.

overlap in the tweets across the keywords. This means, the same tweet can be categorized under two different keywords and it is important to remove the duplicates and not overwhelm the user with redundant information.

5. Conclusion

In this research work, we have demonstrated a personalized recommendation system, based on user profile matching. We have also presented the effectiveness of using a synonym binder for avoiding information loss and enhancing the knowledge extraction process, which was also supported by a sentiment analyzer. Sentiment analysis shows people attitude towards different topics which can be used to generate a richer user profile and personalized recommendations. Topic-based sentiment analysis can generate a rich user profile, personalized recommendation, and helps the user to gather summarized public opinions on entities of their interest. Domain-specific seed words helped to decrease information loss during the keyword-based search. User-generated profile from social media can be integrated with clinical decision support system (CDSS) or electronic health record (EHR) to know about user interest and behavior in detail. In future, we are planning to integrate user information from other social media and user activities log to find interesting patterns and use them in personalized recommender systems.

Data Availability

The data and code related to the data will be made available on Github.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research work was supported by Zayed University Cluster Research Fund, no. R18038.

References

[1] Digital in 2017: Global Overview, We Are Social, 2017, https://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview.

[2] R. Batool, W. A. Khan, M. Hussain et al., "Towards personalized health profiling in social network," in *Proceedings of the 6th International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM)*, IEEE, Taipei, Taiwan, October 2012.

[3] Company Info, Facebook Newsroom, 2019, http://newsroom.fb.com/company-info/.

[4] Most Popular Social Networks Worldwide as of January 2019, Ranked by Number of Active Users (in Millions), 2019, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[5] Q. You, S. Bhatia, and J. Luo, "A picture tells a thousand words-About you! User interest profiling from user generated visual content," *Signal Processing*, vol. 124, pp. 45–53, 2016.

[6] F. Persia and D. D'Auria, "A survey of online social networks: challenges and opportunities," in *Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, pp. 614–620, San Diego, CA, USA, August 2017.

[7] A. V. Lakshmi, S. B. R. Kumar, P. J. Charles et al., "Survey paper on mobile social networks," *International Research Journal of Engineering and Technology*, vol. 2, no. 6, pp. 637–641, 2015.

[8] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PLoS One*, vol. 6, no. 5, Article ID e19467, 2011.

[9] A. Weiler, M. Grossniklaus, M. H. Scholl et al., "Survey and experimental analysis of event detection techniques for twitter," *The Computer Journal*, vol. 60, no. 3, pp. 329–346, 2017.

[10] A. Crisci, V. Grasso, P. Nesi et al., "Predicting TV programme audience by using Twitter based metrics," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 12203–12232, 2018.

[11] S. J. McConnell, "Twitter and the 2016 U. S. presidential campaign: a rhetorical analysis of a rhetorical analysis of tweets and media coverage by Stephen J. Mcconnell," A thesis Submitted in Partial Fulfillment of the Degree of Master of Science in Professional Writing December 2015 New York University School of Professional Studies, New York, NY, USA, 2016.

[12] E. Mohammadi, M. Thelwall, M. Kwasny, and K. L. Holmes, "Academic information on twitter: a user survey," *PLoS One*, vol. 13, no. 5, Article ID e0197265, 2018.

[13] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial," 2015, https://arxiv.org/abs/1505.03105.

[14] B. Al-Jenaibi, "The twitter revolution in the gulf countries," *Journal of Creative Communications*, vol. 11, no. 1, pp. 61–83, 2016.

[15] L. M. Yonker, S. Zan, C. V. Scirica, K. Jethwani, and T. B. Kinane, ""Friending" teens: systematic review of social media in adolescent and young adult health care," *Journal of Medical Internet Research*, vol. 17, no. 1, p. e4, 2015.

[16] Networking Health: Prescriptions for the Internet, 2017, http://www.ncbi.nlm.nih.gov/books/NBK44714/.

[17] S. Fox, "Health topics," 2017, http://www.pewinternet.org/2011/02/01/health-topics-2/.

[18] M. Ybarra and M. Suman, "Reasons, assessments and actions taken: sex and age differences in uses of Internet health information," *Health Education Research*, vol. 23, no. 3, pp. 512–521, 2008.

[19] S. S. Tan and N. Goonawardene, "Internet health information seeking and the patient-physician relationship: a systematic review corresponding author," *Journal of Medical Internet Research*, vol. 19, no. 1, p. e9, 2017.

[20] E. Basch, A. M. Deal, M. G. Kris et al., "Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial," *Journal of Clinical Oncology*, vol. 34, no. 6, pp. 557–565, 2019.

[21] P. H. Keckley and M. Hoffmann, *Social Networks in Health Care: Communication, Collaboration and Insights*, Deloitte Center for Health Solutions, New York, NY, USA, 2010.

[22] Patientslikeme, 2017, https://www.patientslikeme.com/.

[23] A. Sarker, R. Ginn, A. Nikfarjam et al., "Utilizing social media data for pharmacovigilance: a review," *Journal of Biomedical Informatics*, vol. 54, no. 1, pp. 202–212.

[24] V. Ehrenstein, H. Nielsen, A. B. Pedersen, S. P. Johnsen, and L. Pedersen, "Clinical epidemiology in the era of big data: new opportunities, familiar challenges," *Clinical Epidemiology*, vol. 9, pp. 245–250, 2017.

[25] P. Wicks, D. L. Keininger, M. P. Massagli et al., "Perceived benefits of sharing health data between people with epilepsy on an online platform," *Epilepsy & Behavior*, vol. 23, no. 1, pp. 16–23, 2012.

[26] Social Media Likes Healthcare: From Marketing to Social Business, 2017, http://www.pwc.com/us/en/health-industries/publications/health-care-social-media.jhtml.

[27] R. Batool, A. M. Khattak, J. M. Hashmi, and S. Lee, "Precise tweet classification and sentiment analysis," in *Proceedings of the 12th International Conference on Computer and Information Science (ICIS), 2013 IEEE/ACIS*, IEEE, Niigata, Japan, June 2013.

[28] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics—challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, vol. 39, pp. 156–168, 2018.

[29] F. Emmert-Streib, O. P. Yli-Harja, M. Dehmer, and F. Emmert-Streib, "Data analytics applications for streaming data from social media: what to predict?" *Frontiers in Big Data*, vol. 1, p. 2, 2018.

[30] O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez et al., "Fusing pattern discovery and visual analytics approaches in tweet propagation," *Information Fusion*, vol. 46, pp. 91–101, 2018.

[31] S. Petrovic, M. Osborne, V. Lavrenko et al., "RT to win! predicting message propagation in twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, vol. 13, pp. 586–589, Barcelona, Catalonia, Spain, July 2011.

[32] H. Li, D. Caragea, C. Caragea, and N. Herndon, "Disaster response aided by tweet classification with a domain adaptation approach," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 16–27, 2018.

[33] J. Chen, R. Nairn, L. Nelson et al., "Short and tweet: experiments on recommending content from information streams," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, Atlanta, GA, USA, April 2010.

[34] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing user modeling on twitter for personalized news recommendations," in *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, Springer Berlin Heidelberg, Girona, Spain, July 2011.

[35] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of Twitter posts for user profile construction on the social web," in *Proceedings of the Extended Semantic Web Conference*, Springer Berlin Heidelberg, Heraklion, Crete, Greece, May 2011.

[36] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web," in *Proceedings of the 3rd International Web Science Conference*, ACM, Koblenz, Germany, June 2011.

[37] G. Piao and J. G. Breslin, "Exploring dynamics and semantics of user interests for user modeling on twitter for link recommendations," in *Proceedings of the 12th International Conference on Semantic Systems*, ACM, Leipzig, Germany, September 2016.

[38] Z. Deng, M. Yan, J. Sang, and C. Xu, "Twitter is faster: personalized time-aware video recommendation from twitter to YouTube," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2, p. 31, 2015.

[39] I. Celik, F. Abel, and G.-J. Houben, "Learning semantic relationships between entities in twitter," in *Proceedings of the International Conference on Web Engineering*, Springer Berlin Heidelberg, Paphos, Cyprus, June 2011.

[40] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.
[41] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques," in *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, IEEE, Melbourne, FL, USA, December 2003.
[42] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture*, ACM, Sanibel Island, FL, USA, October 2003.
[43] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2007*, pp. 219–222, Boulder, CO, USA, March 2007.
[44] L. Guo and M. Lease, "Personalizing local search with twitter," in *Proceedings of the Workshop on Enriching Information Retrieval at the 34th Annual Association for Computing Machinery Special Interest Group on Information Retrieval Conference*, Beijing, China, 2011.
[45] E. Clark and K. Araki, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English," *Procedia—Social and Behavioral Sciences*, vol. 27, pp. 2–11, 2011.
[46] D. Laniado and M. Peter, "Making sense of twitter," in *Proceedings of the International Semantic Web Conference*, Springer Berlin Heidelberg, Shanghai, China, November 2010.
[47] U. Lösch and D. Müller, "Mapping microblog posts to encyclopedia articles," *Lecture Notes in Informatics*, vol. 192, p. 150, 2011.
[48] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. Welpe, "Predicting elections with twitter: what 140 characters reveal about political sentiment," in *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*, pp. 178–185, Washington, DC, USA, May 2010.
[49] LIWC, 2016, http://www.liwc.net/.
[50] Archivist, 2017, http://archivist.visitmix.com/.
[51] Alchemy API, 2017, http://www.alchemyapi.com.
[52] V. Beal, "twitter dictionary: a guide to understanding twitter lingo," 2017, http://www.webopedia.com/quick_ref/Twitter_Dictionary_Guide.asp.
[53] Java API for WordNet Searching (JAWS), 2017, https://github.com/fcr/JAWS.
[54] Opinion Mining, Sentiment Analysis, and Opinion Spam Detection, 2017, http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html.
[55] American Diabetes Assosiation, 2017, http://www.diabetes.org/diabetes-basics/common-terms/.
[56] Glossary of Diabetes, 2017, https://en.wikipedia.org/wiki/Glossary_of_diabetes.