Healthcare Social Data Platform Based on Linked Data and Machine Learning

Salma El Hajjami, Mohammed Berrada and Soufiane Fhiyil

Abstract The healthcare system is facing very important challenges in order to improve the whole system performance. Different communities are interested in this subject from different perspectives ranging from technical issues to organizational aspects. An important aspect of this research area is to consider social network data within the system especially because of the rapid and growing development of social networks. It can be general social networks, like Facebook or twitter but also others dedicated as PatientsLikeMe. This social network proliferation generates complex problems and locks when we want to take into account the resulting large amounts of data, created continuously, within the healthcare system. We call these data "social data". The aim of this work is to demonstrate that is possible and feasible to build promising alternatives of the traditional healthcare system to improve the quality of services and reduce cost. In our opinion, taking into account "social data" can provide efficient healthcare decisional support systems to help healthcare operators to make optimal and efficient decisions in dynamic and complex environments. Our approach involves data extraction from multiple social networks, data aggregation, and the development of a semantic model in order to answer high-level users' queries. In addition, we show how an analytical tool can help operators to understand data. Lastly, we present a model of machine learning which aims to detect the Sentiments of users expressed toward a given medication and the "TOP TRENDING" of care and treatments used for a given disease.

Keywords Social network · Social data · Semantic model · Analytical tools · Health information · Machine learning · Sentiment top trending

S. El Hajjami (✉) · M. Berrada
Laboratoire D'Informatique et Physique Interdisciplinaire (LIPI), ENS-Fès, USMBA, Fez, Morocco
e-mail: salma.elhajjami@usmba.ac.ma

S. Fhiyil
Faculté des Sciences Dhar El Mehraz, USMBA, Fez, Morocco

© Springer Nature Singapore Pte Ltd. 2020

V. Bhateja et al. (eds.), *Embedded Systems and Artificial Intelligence*,
Advances in Intelligent Systems and Computing 1076,
https://doi.org/10.1007/978-981-15-0947-6_28

salma.elhajjami@usmba.ac.ma

1 Introduction

Today, we are witnessing a veritable deluge of data produced by social media. This data is generated from a large number of Internet applications and Web sites. Facebook is the most popular platform (with more than 1.19 billion active users per month), followed by Twitter (500 million users worldwide). In parallel of general use social networks, we also find platforms dedicated to health, such as Doctissimo (eight million unique visitors each month) [1] and PatientsLikeMe (growing, currently has more than 187,000 members and covers more than 500 patients) [2]. These platforms are examples of medical social networking sites with many user bases, where people with specific diseases can exchange information about their illness, treatment, and experience.

This profusion of digital traces left by users of social media is at the origin of a new phenomenon very popular in recent years, and it is about social data. This phenomenon profoundly transforms many sectors, like health. In this field, the available data represent an unprecedented innovation potential: identification of disease risk factors, assistance with diagnosis, choice and monitoring of treatment effectiveness, pharmacovigilance, epidemiology, etc. In this sense, taking into account "social data" makes it possible to respond to market needs and trends in healthcare institutions [3], and to offer epidemiologists, physicians, and health policy experts an excellent opportunity to formulate evidence-based judgments that will eventually lead to patient care [4].

However, everyone agrees that the treatment of these masses of data, or "social data", is a major issue [5]. Collecting, processing, and analyzing large social media data from unstructured (or semi-structured) sources to extract valuable knowledge are an extremely difficult task that has not been fully resolved. Traditional data management methods, algorithms, frameworks, and tools have become inadequate to handle this large amount of data. This problem has generated a large number of challenges related to different aspects such as knowledge representation, data collection, processing, analysis, and visualization [5]. Some of these challenges include access to very large amounts of unstructured data (management problems), determining the amount of data needed to obtain a large amount of high-quality data (quality vs. quantity), and the processing of the data stream changing dynamically.

However, given the very large number of heterogeneous data from social media, one of the main challenges is to identify and analyze the useful data in order to discover useful knowledge to improve the decision making of individual users and enterprises [6]. As such, the integration and analysis of these data, key elements of patient privacy, and vital tools for healthcare professionals are not a trivial task due to their scalability, complexity, and heterogeneity. Traditional analysis techniques and methods (data analysis) need to be adapted and integrated with the new data paradigms that have emerged for mass data processing.

The main objective of our work is to develop a platform that allows integrating and visualizing knowledge from a large amount of information available on "patientslikeme.com", with the inclusion of "Twitter". We use linked data technologies

salma.elhajjami@usmba.ac.ma

to aggregate all this data into a semantic model in order to arrive at contextually relevant information and generate knowledge. Thus, we present a model of automatic learning which aims to detect the "TOP TRENDING" drugs used for a given disease. The platform also supports a set of analytical tools that help the user to become aware of socially distributed health information. The paper is organized as follows. Section 2 surveys the related work on healthcare data integration that identifies the research issues involved. In Sect. 3, we present our semantic approach that we used for integrating health knowledge, and in Sect. 4, the visualization results are illustrated, interpreted, and discussed, followed by conclusions and future work in Sect. 5.

2 Related Work

This section discusses the state of the art in the area of integration and analysis of social data available on social networks.

Social media has been integrated into medical practice and has reshaped healthcare services in several ways. It has been proven a viable platform for patients to discuss health-related issues [7] and for researchers to derive health intelligence [8]. However, integrating data from the social web is a challenging task. Several works surveyed approaches to extracting information from the "social web" for health personalization. Research works [9, 10] extract health information from different sources, including the web and social media, sensors, healthcare claims and laboratory images, and physician notes that provide useful health information. The authors of [11] used social media to effectively build novel disease surveillance systems that detect, track, and respond to infectious diseases, such as in the case of the 2009 H1N1 Influenza. Social media has improved healthcare quality with better communication between patients and clinicians, either through generic channels such as Facebook or Twitter [12], or via special sites such as PatientsLikeMe for patients.

Recently, the use of SW technology as a framework for the integration of public data has become popular. Most of the work in this thread follows Linked Open Data (LOD) [13, 14] principles to create links between resources distributed in heterogeneous data sources. In [15, 16], an OWL ontology is used to link the schemas of semi-structured and structured data. In the field of health, a semantic integration model of different health data sources that can help annotate social health blogs is used [17]. The work [18] uses an integrated semantic model to create a machine-readable encoding of the content semantics of various open health data sources, especially social data sources.

Even though there are many sentiment analyses of Tweets in general area [19, 20] and in the health domain [21] using data mining and machine learning approach, most of the works do not apply the results of the sentiment analysis to measure the degree of public concerns or anxiety toward disease. Data mining and machine learning techniques are used to predict disease risks for individuals or to rank diseases by

salma.elhajjami@usmba.ac.ma

their risks. For instance, in [22, 23], a condition for one patient is predicted using similar patients, based on 13+ million elderly patients' hospital visit records.

3 Semantic Approach for Health Knowledge Integration

In this section, we present our approach that we implemented in Python. We describe in detail the methods and techniques used to extract data from social networks on the Web. We also present a "Matching-Term" algorithm that aims to validate the "tweets" extracted by calculating the similarity between the term used by the user in his "tweet" and the list of medical terminologies extracted from the "UMLS" database [24]. Finally, the "Mapping" of relational model to "RDF" semantic model is described.

3.1 Data Collection

The collection of data from social networks is a big part of our project, and it is the most important phase. In this phase (Fig. 1), we start with the social network "PatientsLikeMe"; this site contains information on patient profiles such as their personal details "Surname, first name, phone …", as well as information on different diseases and their treatments "drugs". We will then extract the "tweets" of users who speak about a given treatment; these treatments are already extracted from "PatientsLikeMe"; and therefore, for each treatment, we must extract a number of "tweets" to find out what are the most popular treatments for a given disease (TOP TRENDING) and also to detect the feelings of users regarding a treatment.

Fig. 1 Data collection

salma.elhajjami@usmba.ac.ma

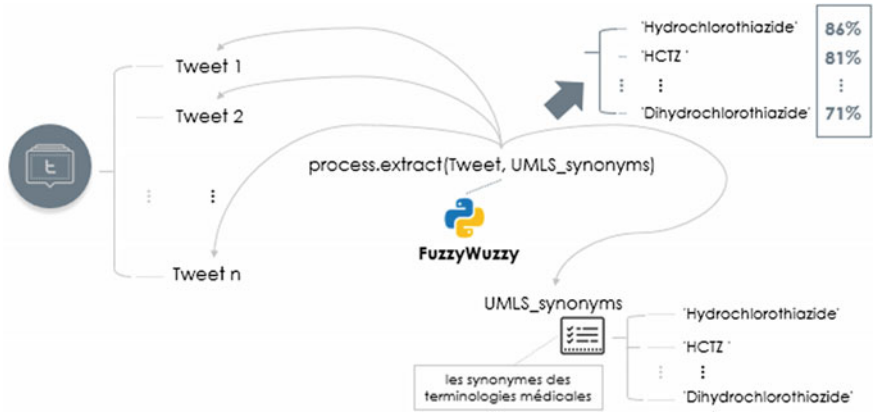

Fig. 2 Cleaning data

3.2 Cleaning Data

After collecting "tweets", we must apply a cleanup before storing them. We use the UMLS [24] Metathesaurus, which provides a common vocabulary and semantics for multiple terms, that refer to the same concept. In this paper, we utilize a "Matching-Term" algorithm (Fig. 3), using the UMLS to recognize identical concepts. This relational database occupies more than "30 GB" of space and contains over 300 million lines. We also use a package called "FuzzyWuzzy" [25] to find the similarity

Algorithm 1: Matching Term

Input : tweets, dict_synonyms, threshold

```
Browse each "tweet" from the "tweets" input dataset ;
Search in the medical dictionary "dict_synonyms" the list of
  terminologies corresponds to the term (i.e. name of the treatment) used
  in the "tweet";
Get the different synonyms terms and add the terms in a list
  "[UMLSTreatement_Terms]";
Use the "FuzzyWuzzy" library which calculates the similarity between a
  list of entries and a string of characters which is in our list of medical
  synonyms "[UMLSTreatement_Terms]" and the "tweet";
Initialize the filter rate: (set thresold at 70%);
foreach ForEach similarity_ratio in FuzzyWuzzy,
  extract([UMLSTreatement_Terms] , tweet): do
    if similarity_ratio > threshold then
        store_Tweet_FileCSV ();
        break;
    end
end
```

Fig. 3 "Matching-Term" algorithm

salma.elhajjami@usmba.ac.ma

between the medical synonyms and the medical term that the user used in his "tweet" which has among its features a comparison function of a character string with a list of character strings. However, we give this function in the parameters the collected "Tweet" and the list of medical synonyms that we extracted from "UMLS", and so, if we find that the medical term used in the "Tweet" is similar with at least a medical term in the list of synonyms more than the degree of threading that we set (e.g., 70%), then we will keep the "Tweet" and persist it in an Excel file "CSV" (Fig. 2).

The purpose of this algorithm is to validate the "tweets"; it means that when a surfer on "Twitter" speaking on a given treatment, we must be sure that the medical term (name of treatment) that he used in his "tweet" complies with at least one terminology in the list of synonyms of this treatment.

3.3 Applying Machine Learning to Detect Sentiment

Now that we collected all tweets and cleaned them by "Matching-Term" algorithm, we present the best model constructed on a training dataset, it is called "Sentiment140" [26], and it has been developed by a group of student researchers at University of Stanford, and the way they created this dataset is that any tweet with positive emotions, like :), were positive, and tweets with negative "emotions", like :(, were negative. It contains more than 1.6 million records, and data are annotated in two classes "positive, negative". Therefore, we used this data frame to train our models, and we selected the best model after a set of tuning parameters of every algorithm and also applying dimensionality reduction in order to reduce features, but this last did not give better results, so at the end we did not use dimensionality reduction approach. We tested two approaches in order to present textual data into numerical data "Count Vectorizer" and "TF-IDF Vectorizer" [27], but the one that gave better results in terms of accuracy performance is "TF-IDF Vectorizer". We evaluated our models in a cross-validation of fivefolds, and we used accuracy metric to measure their performance. As we can see in (Fig. 4), we have three best performing models "LinearSVC", "Logistic Regression", "Ridge Classifier" but the best one is logistic regression with 0.82 accuracy score evaluated with cross-validation. Eventually, we used logistic regression model to predict new tweets talking about different treatments in order to detect their sentiments.

3.4 Transformation to RDF

To perform intelligent analyzes across multiple data sources, we use a lightweight ontology to build an integrated knowledge base. Using the entities in the data model, we developed the lightweight ontology by structuring entities and relationships ontologically as a concept hierarchy as shown in Fig. 5.

salma.elhajjami@usmba.ac.ma

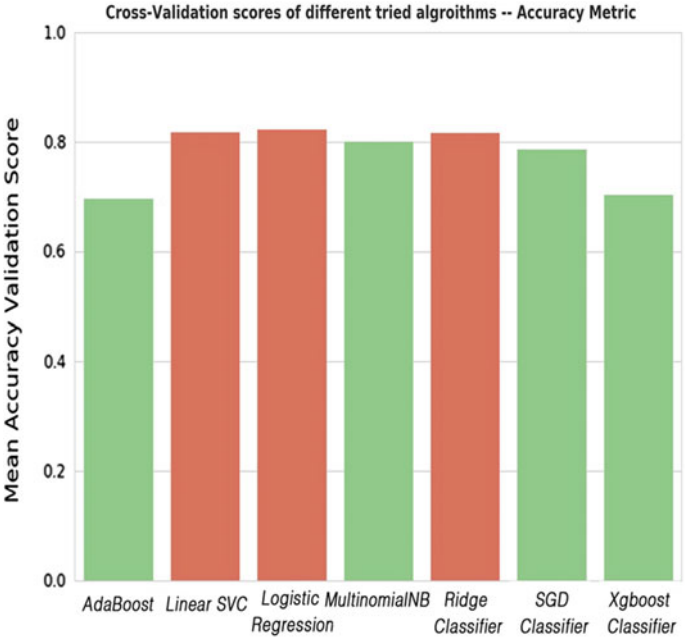

Fig. 4 Cross-validation

An ontology typically consists of classes, properties, relationships, between classes (e.g., IS-A, PART-OF), and instances.

To capture the variety of health information, we developed a conceptual model to capture the relationships among health data entities extracted from different sources. A patient can have a certain disease; this is modeled with the "Patient" class, which has a predicate relation called "hasCondition" with the "Condition" class. A disease can have symptoms and treated with several treatments; this is modeled with the "Condition" class, which has a predicate relation called "show" with the "Symptom" class and another predicate relation called "treatedBy" with "Treatment" class. However, each of the classes "Side-Effects" and "Symptom" have severity levels and their number evaluations, this is modeled through the nodes "EvaluationSeveritySymtpom" and "EvaluationSeverityTreatment" and their properties "hasSeverity", "hasEvaluation". Finally, we model the "tweets" we extracted for each treatment, through the "Treatment" class that has a predicate relation called "hasTweets" with the "Tweets" class. The "Tweets" class has a predicate relation called "hasTweetDate" with the "TweetDate" class which models the fact that a tweet has a own date and also his own feeling.

salma.elhajjami@usmba.ac.ma

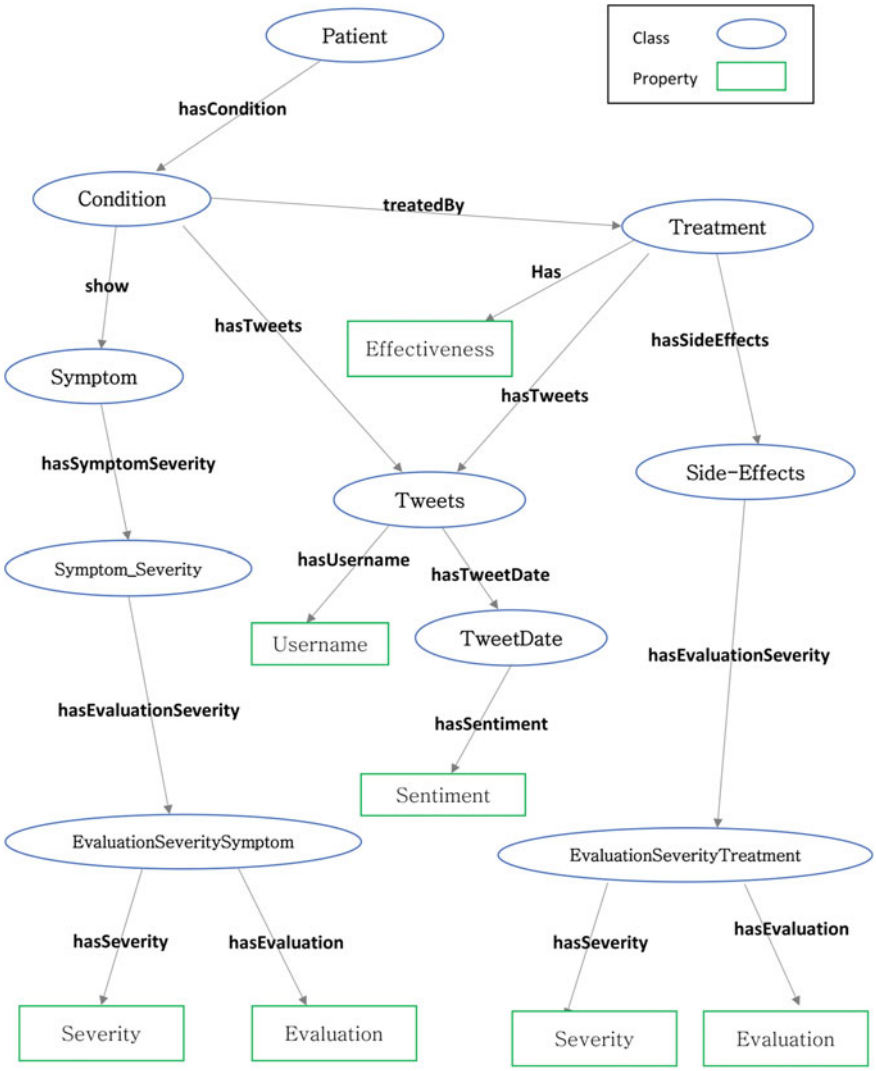

Fig. 5 Semantic model for social health entities

4 Implementation and Results

In this section, we present our platform with various analytical features allowing the extraction of data from social networks "PatientsLikeMe, Twitter", the detection of the most popular treatments TOP TRENDING used by patients and finally the prediction of the sentiments expressed by the patients who have followed these treatments (Fig. 6).

salma.elhajjami@usmba.ac.ma

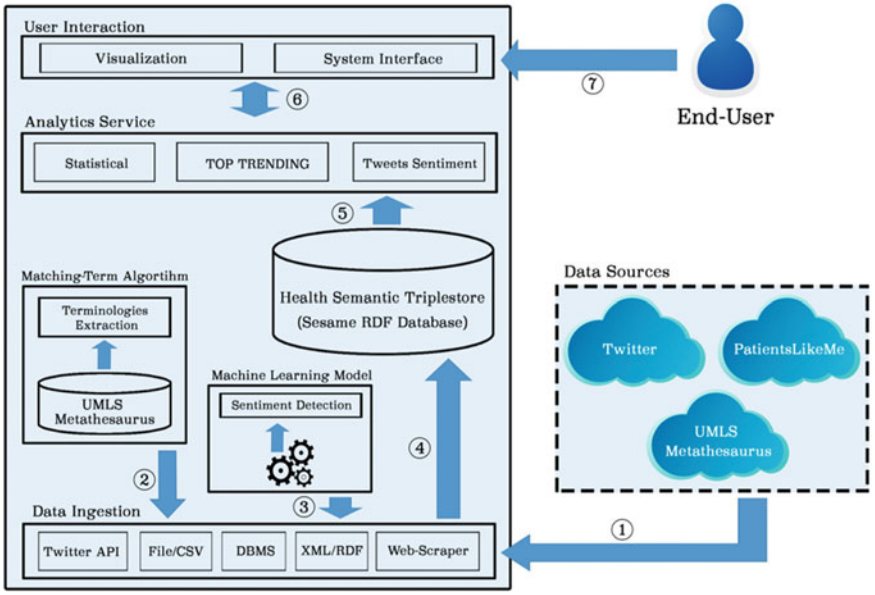

Fig. 6 Platform architecture

4.1 Platform Architecture

The figure above shows the architecture of our platform with all steps taken:

1. On the lower level of the architecture, we have the data ingestion layer. This layer is responsible for the extraction of data from various sources of data on health available to the public. The layer is composed of several connectors, one for each type of data source. Most of the Web sites of health do not provide an API that allows researchers to retrieve data. In addition, we used a Web Scraper called "Scrapy" in order to target Web sites and extract relevant information. We have used data sources like PatientsLikeMe, UMLS, and Twitter (via APIs).
2. After the extraction of data from Twitter's social network, we must also apply our "Matching-Term" algorithm that is designed to validate the extracted "tweets" by comparing the medical term used by the user in his "tweet" and the list of medical synonym terms extracted through the UMLS database, and so, if there is at least one term in the list similar to the term used by the user, then we keep the "tweet" by saving it in an Excel file.
3. In this layer, after having validated all the tweets by our algorithm "Matching-Term", we must now use our machine learning model to predict the sentiment of all tweets and save them into our relational database MySQL.
4. In this layer, we migrate toward the semantic model RDF that we built, we will apply our Python script which allows the mapping from the relational model

salma.elhajjami@usmba.ac.ma

to the semantic model, and then, we will save the converted data into the RDF server called "Sesame" that allows you to store RDF data [28].
5. In this layer, we present the analytical tools that our platform will support such as the detection of patient users' sentiments toward the treatments they took, and the identification of the most popular TOP TRENDING treatments used for a given disease.
6. Last but not least, we need just to present these tools in different types of visualizations on our platform such as the "Pie charts …".
7. The last layer, users can interact now with our platform via the visualizations or the interface system, which invokes analysis operations based on the request sent by the user.

4.2 Platform Analytics Tools

In this section, we present the interface of the platform, which is in the form of a dashboard containing different visualizations and types of graphics.

4.2.1 Presentation of the Most Popular Treatments

In order to detect the most popular treatments, we must construct a SPARQL query that is designed to query RDF data from the "Sesame" server and then return the results needed to be displayed to the user. Therefore, the patient can select a certain disease; secondly, he should mention the start date and the end date, because popularity is something relative to time. For example, if a treatment is popular in a certain period of time, perhaps it will be more popular in another time, so the date is very important in the ranking of popularity.

Figure 7 shows the most popular treatments according to the disease that the patient has selected; in this case, it is "type 1 Diabetes". The message shows that 922,737 users on Twitter talk around these treatments, and the data are based from 2018-03-06 to 2018-03-28.

4.2.2 Presentation of Patient Data

Figure 8 shows the locations of patients who are diagnosed with the type 1 diabetes; these data came from the profiles of patients that we have extracted from "Patientslikeme".

salma.elhajjami@usmba.ac.ma

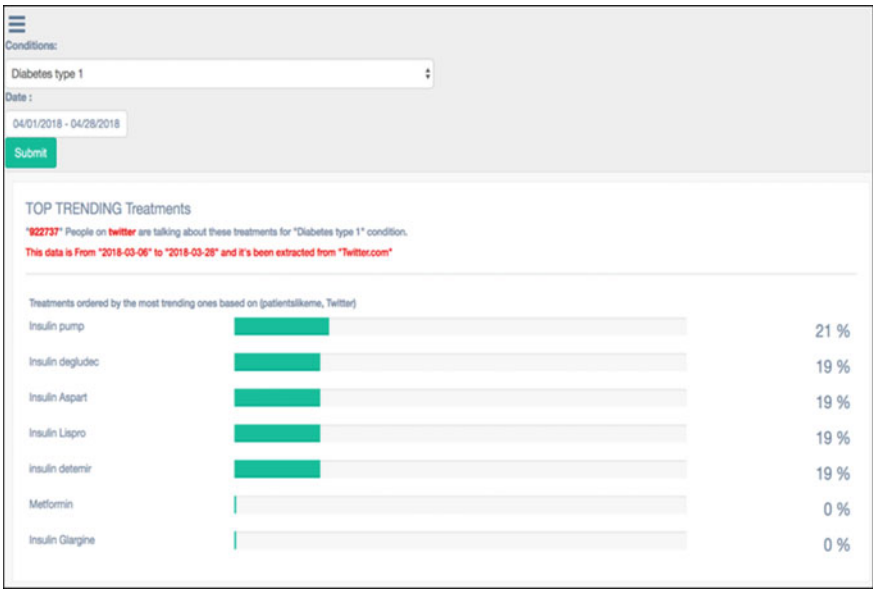

Fig. 7 TOP TRENDING treatments

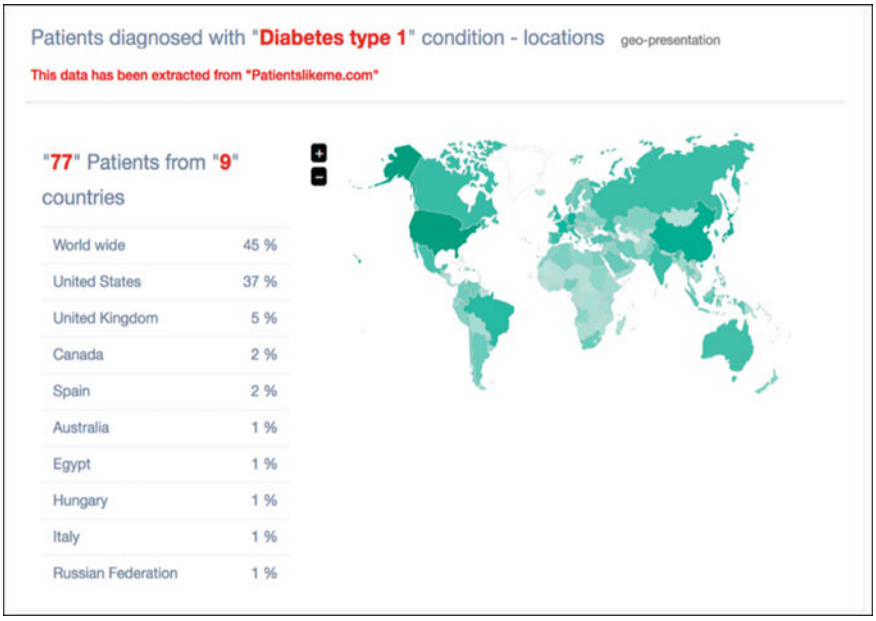

Fig. 8 Patients diagnosed with a given disease

salma.elhajjami@usmba.ac.ma

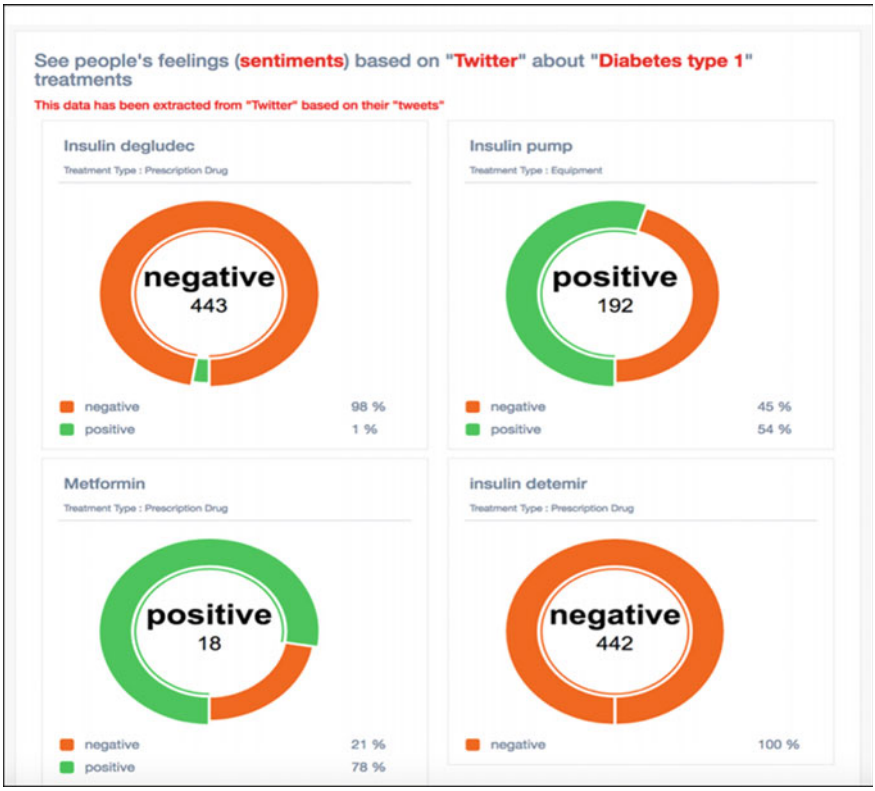

Fig. 9 Feelings of "Twitter" about treatment

4.2.3 Presentation of Sentiments Expressed by Users in Their Tweets

In this section, we also present graphical visualizations about the feelings expressed by users in their "tweets" about a certain treatment (Fig. 9).

The figure above shows the graphical visualizations "Pie Charts" concerning the people's feelings about different treatments. This part was realized using the machine learning model.

5 Conclusion and Perspectives

The field of health is a very complex and dynamic area where decision makers often rely on decision support systems to analyze and compare the actors involved (clinical trials, procedures, etc.). The expansion of the social web in the health field increases considerably the flow of data, where the experiences of patients are shared such as treatments, side effects, etc. These data can be important sources to provide collective

salma.elhajjami@usmba.ac.ma

intelligence, evaluate, and improve healthcare performance. As a result, the goal of our work was to develop a platform that synthesizes knowledge from a large body of data extracted from "PatientsLikeMe, Twitter" social networks, in order to identify the most used treatments by the patients and also to detect their feelings expressed toward these treatments.

In our future work, we aim to use several sources of medical data from many social networks, such as "Blogs, Forums, Datasets …" which contain, for example, information about patients' discussions, their questions asked in the forums, and the different health data sets that have been published by the government. There are also other sources of data that contain a lot of information such as articles and scientific journals of health. We plan after exploring all these data sources to build a semantic model that will be able to group all of these data together to generate knowledge. We also aim to implement these results in an extended platform that will contain numerous queries and analytical tools to better serve policy makers and patients.

References

1. Doctissimo. http://www.doctissimo.fr/
2. PatientsLikeMe. https://www.patientslikeme.com/
3. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Heal. Inf. Sci. Syst. **2**, 1–10 (2014). https://doi.org/10.1186/2047-2501-2-3
4. Sessler, D.I.: Big data and its contributions to peri-operative medicine. Anaesthesia **69**, 100–105 (2014)
5. Kaisler, S., Armour, F., Espinosa, J.A., Money, W.: Big data: issues and challenges moving forward. In: Proceedings of 46th Hawaii International Conference on System Sciences (HICSS), pp. 995–1004. IEEE (2013)
6. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact, MISQ **36**(4), 1165–1188 (2012)
7. Househ, M., Borycki, E., Kushniruk, A.: Empowering patients through social media: the benefits and challenges. Health Inf. J. **20**, 50–58 (2014)
8. Ji X, Chun SA, Geller J.: Monitoring public health concerns using Twitter sentiment classifications. In: Proceedings of IEEE International Conference on Healthcare Informatics, pp. 335–344. IEEE, Philadelphia, PA (2013)
9. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**, 1–10 (2014)
10. Ji, X., Chun, S.A., Geller, J.: Monitoring public health concerns using Twitter sentiment classifications. In: Proceedings of IEEE International Conference on Healthcare Informatics, pp. 335–344. Philadelphia, PA (2013)
11. Brownstein, J.S., Freifeld, C.C., Chan, E.H., Keller, M., Sonricker, A.L., Mekaru, S.R., Buckeridge, D.L.: Information technology and global surveillance of cases of 2009 H1N1 influenza. N. Engl. J. Med. **362**(18), 1731–1735 (2010)
12. Bull, S.S., Breslin, L.T., Wright, E.E., Black, S.R., Levine, D., Santelli, J.S.: Case study: an ethics case study of HIV prevention research on Facebook: the just/us study. J. Pediatr. Psychol. **36**(10), 1082–1092 (2011)
13. Bizer, C.: Evolving the Web into a Global Data Space. In: Fernandes, A.A., Gray, A.G., Belhajjame, K. (eds.). Proceedings of 28th British National Conference on Databases, p. 1. Springer Berlin Heidelberg, Manchester (2011)
14. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. Int. J. Semant. Web Inf. Syst. **5**, 1–22 (2009)

salma.elhajjami@usmba.ac.ma

15. Skoutas, D., Simitsis, A.: Designing ETL processes using semantic web technologies. In: DOLAP, pp. 67–74 (2006)
16. Skoutas, D., Simitsis, A.: Ontology-based conceptual design of ETL processes for both structured and semi-structured data. IJSWIS **3**(4), 1–24 (2007)
17. Chun, S.A., Mac Kellar, B.: Social health data integration using semantic web. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, pp. 392–397 (2012)
18. Ji, X., et al.: Linking and using social media data for enhancing public health analytics. J. Inf. Sci. **43.2**, 221–245 (2017)
19. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008) Social Health Records: Gaining Insights into Public Health … 41
20. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 43–50. Arlington, VAS (2006)
21. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS ONE **5**(11), e14118 (2010)
22. Chawla, N.V., Davis, D.A.: Bringing big data to personalized healthcare: a patient-centered framework. J. Gen. Intern. Med. **28**, 660–665 (2013)
23. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabasi, A.L.: Time to CARE: a collaborative engine for practical disease prediction. Data Min. Knowl. Disc. **20**, 388–415 (2010)
24. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(suppl_1), D267–D270 (2004)
25. Gonzalez, J.: fuzzywuzzy Fuzzy String Matching in python. https://github.com/seatgeek/fuzzywuzzy
26. http://help.sentiment140.com/for-students
27. Soucy, P., Mimeau, G.W.: Beyond TF-IDF weighting for text categorization in the vector space model. In: Proceedings of 19th International Joint Conference Artificial Intelligence (IJCAI '05), pp. 1130–1135 (2005)
28. Broekstra, J., Kampman, A., Van Harmelen, F.: Sesame: an architecture for storing and querying RDF data and schema information (2001)

salma.elhajjami@usmba.ac.ma