Automatic extraction and identification of users' responses in Facebook medical quizzes

Alejandro Rodríguez-González[a,*,1], *Ernestina Menasalvas Ruiz*[b,1], *Miguel A. Mayer Pujadas*[c,1]

[a] *ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte, 28660 Madrid, Spain*
[b] *Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Campus de Montegancedo, Pozuelo de Alarcon, 28223 Madrid, Spain*
[c] *Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute (IMIM) – Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona (PRBB), Dr. Aiguader, 8808003 Barcelona, Spain*

A B S T R A C T

Background: In the last few years the use of social media in medicine has grown exponentially, providing a new area of research based on the analysis and use of Web 2.0 capabilities. In addition, the use of social media in medical education is a subject of particular interest which has been addressed in several studies. One example of this application is the medical quizzes of The New England Journal of Medicine (NEJM) that regularly publishes a set of questions through their Facebook timeline.
Objective: We present an approach for the automatic extraction of medical quizzes and their associated answers on a Facebook platform by means of a set of computer-based methods and algorithms.
Methods: We have developed a tool for the extraction and analysis of medical quizzes stored on Facebook timeline at the NEJM Facebook page, based on a set of computer-based methods and algorithms using Java. The system is divided into two main modules: Crawler and Data retrieval.
Results: The system was launched on December 31, 2014 and crawled through a total of 3004 valid posts and 200,081 valid comments. The first post was dated on July 23, 2009 and the last one on December 30, 2014. 285 quizzes were analyzed with 32,780 different users providing answers to the aforementioned quizzes. Of the 285 quizzes, patterns were found in 261 (91.58%). From these 261 quizzes where trends were found, we saw that users follow trends of incorrect answers in 13 quizzes and trends of correct answers in 248. *Conclusions:* This tool is capable of automatically identifying the correct and wrong answers to a quiz provided on Facebook posts in a text format to a quiz, with a small rate of false negative cases and this approach could be applicable to the extraction and analysis of other sources after including some adaptations of the information on the Internet.

* *Corresponding author.*
E-mail addresses: alejandro.rodriguezg@upm.es (A. Rodríguez-González), ernestina.menasalvas@upm.es (E. Menasalvas Ruiz), mmayer@imim.es (M.A. Mayer Pujadas).
[1] These authors contributed equally to this work.

1. Introduction

In the last few years the use of social media in medicine has grown exponentially, providing a new area of research based on the analysis and use of Web 2.0 capabilities. This has affected the way that people share and exchange ideas, opinions and feelings [1]. Furthermore, it has increased the amount of information available on the Internet under the concept of user generated content (UGC) [2]. In medicine, social media has a deep impact because the number of websites with health-related information has been growing very fast in the last few years [3], making this information available to a wider audience access. However, the lack of control and quality [4] regarding this information also contributes to the generation of low quality medical information which could be dangerous for the potential users [4,5].

In addition, the use of social media in medical education is a subject of particular interest which has been addressed in several studies such as the one undertaken by the Penn State College of Medicine in relation with the use of Twitter, YouTube, Flickr, blogging and Skype to promote students learning [6]. Another interesting analysis was made by Bahner et al. [7] which studies the effectiveness of using Twitter and Facebook as an educational tool by sharing messages related to a particular subject. Cheston et al. [8] also carried out a study where the use of social media in education was analyzed.

In the context of medical education, there are several social media platforms that offer medical knowledge in different ways and formats. Twitter accounts like USMLE [9] or Crush USMLE [10] offer relevant information regarding the United States Medical Licensing Examination (USMLE). Radiology Signs [11] and Radiopaedia [12] are two Facebook pages focused either on the publication of medical images with relevant information or on asking the page users for a diagnosis on the comments of the post. Similar to those previously mentioned is the New England Journal of Medicine (NEJM) Image Challenge Application [13], where your answer to a concrete challenge can be submitted offering basic statistics compared with other answers and interactive medical cases provided by the NEJM official page [14].

Medical quizzes are another interesting category of educational content that can be found in social media platforms. Facebook pages such as Medical Quiz [15] or medical quizzes [16] provide these types of educational questions. However, the format used in these pages is heterogeneous, hindering later analysis of the questions and answers published.

Medical quizzes of The New England Journal of Medicine (NEJM) (one of the most relevant journals in medicine around the world), regularly publishes a set of questions through their Facebook timeline allowing NEJM Facebook page users to answer, discuss or just learn about the knowledge around the question. In particular, the "Medical Quiz" consists of a question about a case study that is accompanied by a set of possible answers. The community manager creates a post where the question is formulated generally with a link to some article or interesting text related to the proposed question. Fig. 1 shows a screenshot with an example of this type of quiz and some answers.

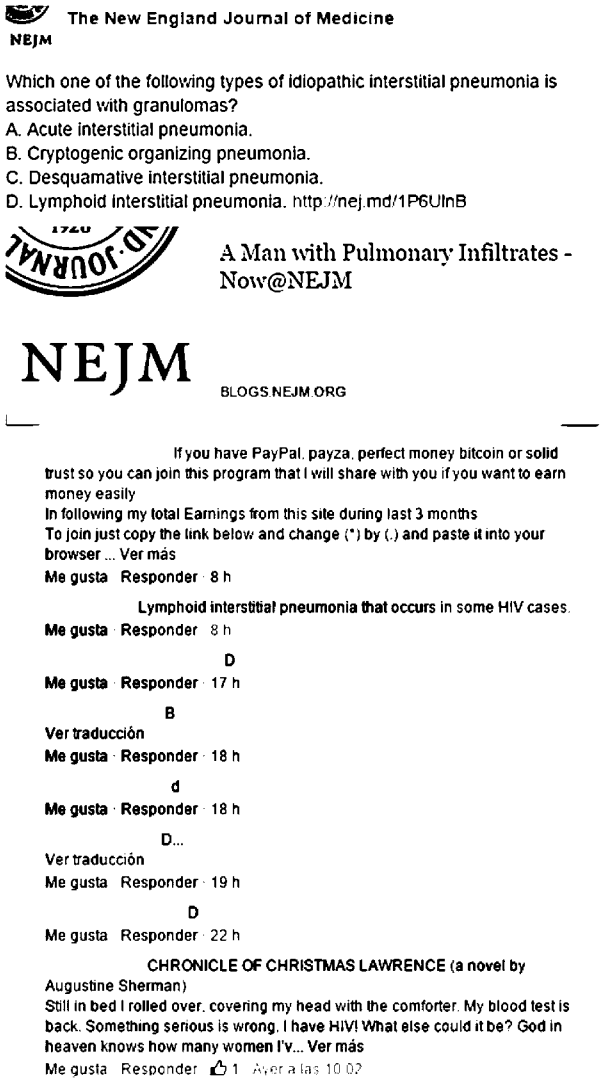

Fig. 1 – Screenshot of a NEJM quiz example.

Once the question is published, users can contribute through their comments. Most of the users try to answer the quiz instead of starting a discussion about the question as the comment analysis confirmed. After a few days, NEJM posts in its timeline the answer to the quiz in a new post.

The benefit behind the effort of NEJM maintaining these quizzes is twofold. On the one hand, it improves the medical education of those users participating and interacting trying to find a correct answer and reading the complementary resources. On the other hand, the comments provided by the different users can be analyzed and mined to extract knowledge hidden in the collective opinion of the users (wisdom of the crowd). The extraction of opinions, feelings or interests is a frequent topic of research in several areas [17–19], used as a basis for developing projects, tools and other efforts oriented to the automatic data extraction from social media [20–23].

Also, the fact that most medical exams around the world are based on medical quizzes increases the interest of the analysis of these kinds of social media platforms quizzes based on intelligent methods.

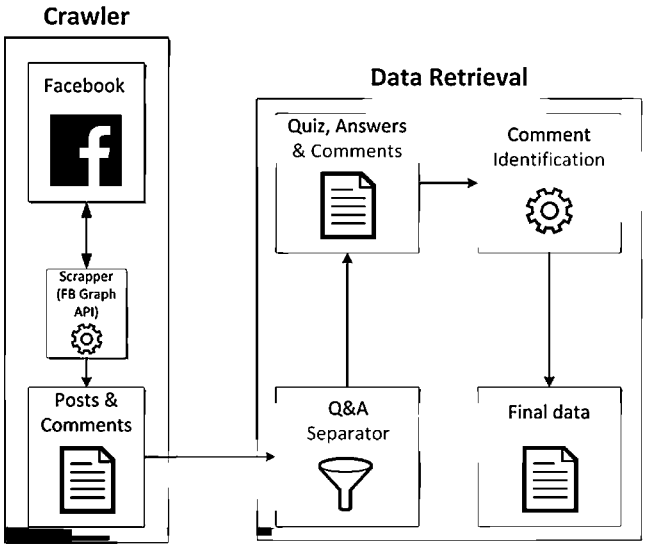

Fig. 2 – Architecture of the approach.

In this paper, we present an approach for the automatic extraction of medical quizzes and their associated answers on a Facebook timeline by means of text processing techniques with the aim of analyzing users' answers. To fulfill the goal we have developed a set of computer-based methods that allows the crawling, processing and identification of medical quizzes and their respective answers together with the comments provided by the users in NEJM Facebook webpage related to these quizzes. Besides, a set of intelligent heuristics that are based on information retrieval from texts have been implemented for the analysis of the extracted information.

The paper is organized as follows: Section 2 presents the methods used in the development of our approach, including the architecture and explanation of the modules involved. Section 3 discusses on the results. Finally, conclusions and future works are presented in Section 4.

2. Methods

Our approach is based on the development of a set of computer-based methods and algorithms for the extraction and analysis of medical quizzes stored on any Facebook timeline (but applied to NEJM Facebook page). The final system is developed entirely in Java (source code of the approach as well as obtained raw data is publicly available online [24]) and is divided in two main modules: Crawler (left) and Data retrieval (right) whose architecture is depicted in Fig. 2.

2.1. *Crawler*

The Crawler is in charge of obtaining all the posts of the Facebook timeline. We created an empty application via Facebook Developers [25] in order to obtain an access token which allows us to use Graph API to crawl the data from the Facebook page [26]. The data produced by the Crawler is saved in a Properties [27] format file and sent to the data retrieval module.

2.2. *Data retrieval (DR)*

This component is responsible for the analysis of the posts retrieved by the Crawler by means of the following steps:

1. Find quizzes and answers to quizzes in the posts.
2. Pair quizzes posts with their corresponding answer posts.
3. Perform the process of identifying correct answers in the answer posts.

Analyze the comments provided by the users in the quiz post to identify the answers posted (if applicable).

Data retrieval: Step 1: *Quiz and answer finding process*
The first step of this workflow is done by the Quiz and Answer (Q&A) separator. The algorithm behind this module works with the high degree of homogeneity existing between the different quizzes and answers. It assumes (after a manual and visual analysis of several NEJM posts which contains quizzes) that a quiz is posted following always a concrete format what makes possible the extraction of the information by means of regular expressions.
Data retrieval: Step 2: *Q&A pairing*
The pairing between questions (quizzes) and answers has been done automatically through the URL. Each posted quiz contains a URL which contains more information about the quiz itself. As the posts with the answer to a quiz also contain the URL (taking into account that the answer is a copy of the question with an additional string identifying the correct answer option) the process of matching is as follows:
a. Find the associated answer to a pre-identified quiz.
b. Find the associated quiz to a pre-identified answer.
Data retrieval: Step 3: *Identification of correct answer in answers posts*
Once the mapping between questions and answers has been performed, a process for identifying the correct answer in a post is carried out. This process analyzes the answer post searching for string patterns that match the identification of a correct answer. The homogeneity of the strings used by the community manager of the journal to identify the correct answer allows to easily finding the correct answer by means of searching for predefined patterns in the text.
Step 4: *Comment analysis*
Once the paired quiz-answer has been generated, the comments provided in the quiz post of the pair are analyzed. This analysis has been performed applying a set of strategies which includes direct matching, Levenshtein distance, tokenization of white spaces, answer found in text and brute force. The process analyzes the comment of a user and the answer provided by NEJM to check whether the comment is the correct answer for the current quiz or alternatively discards the comment as irrelevant (without containing an answer). The output of this final step is stored in a TSV file format.

A description of the different implemented strategies is outlined. Some of these strategies are executed after a string pre-processing which includes the removal of stop words, symbols and convert the string to upper case. It should be

noted that, for the precision analysis of the strategies, two options could be taken into account: when it identifies a correct answer and an incorrect answer. For a better identification of each case we attach to the description of the strategy a code that identifies each case (strategy + correct/incorrect identification).

- *Direct matching (1XX codes)*: three strategies have been developed to find a direct matching between the full comment and the answer. The three strategies respectively detect a direct matching whenever: (a) full comment equals answer option (A, B, C, D...) [1X1] (correct [101] or incorrect [111]), (b) full comment equals answer text [1X2] (correct [102] or incorrect [112]) or (c) full comment equals answer option after the aforementioned preprocessing [1X3] (correct [103] or incorrect [113]).
- *Answer found in text (2X1 codes)*: Unlike the previous case, this strategy is in charge of finding a correct [201] (or incorrect [211]) answer in the comment (not direct equals) ensuring that there is no reference to an incorrect (or correct) answer.
- *Tokenization of white spaces (3X1 codes)*: This strategy has been developed to find a correct [301] (or incorrect [311]) answer using this method. The idea behind is that several users put their answer in the comment after a bunch of text. We use this method to determine if the user did that and provides a correct answer. For example: a user wrote "I consider that correct answer is B because the value of...". After the tokenization of the string using white space as delimiter, we can get this "B". All the tokens obtained are pre-processed removing external symbols so we have a clean B that could be the answer provided by the user, which will be identified as "correct" or "incorrect" answer.

This method has two main problems: Firstly, the use of preposition "a". If the user has used this preposition in upper case it will take it as an answer when it is not. Secondly, the user can make reference to an answer option (B, C, etc.) but they are not stating that this is the correct answer (maybe they are just talking about the option).

However, without a deep analysis using more complex natural language processing algorithms, this is the best option that we have found. We must also stress that a first estimation based on a manual analysis of the comments lead us to think that this is a good solution.

- *Levenshtein (4XX codes)*: two strategies are used to find a correct or incorrect answer using Levenshtein distance [A2]. The first one (4X1), similar to the direct matching, tries to see if the answer provided by the user is correct [401] or not [411] (the entire comment) using Levenshtein distance. The second strategy (4X2) is a combination between Levenshtein and brute force. Imagine a comment like this: "*I consider that the correct answer is Ortostatc hpotnsion because a change of the position of*...". This comment clearly contains a valid answer (Orthostatic hypotension) but has been misspelt. The previous strategies cannot detect this case because the answer is among other words (the main Levenshtein algorithm will not find it) and because it is misspelt (correct answer found in text will not found it). Hence, this strategy creates all the possible combinations of sentences inside a sentence using white space as separator (I, I consider, I consider that,..., consider, consider that, consider that the,..., that, that the, that the correct,...) and check each combination to find a correct [402] (or incorrect [412]) answer by means of the Levenshtein distance. This strategy has the same drawback as white spaces in that the context of the sentence cannot be known, and in some cases may be incorrectly identified.
- *Equals by brute force (Code 5X1)*: similar to the brute force described in the Levenshtein strategy, we try to found a direct matching (full equals) of the different parts created by the brute force strategy. If we found a match, this match can be a correct option [501] or an incorrect one [511].

These strategies are applied over the text comments to find an answer in the comment. The strategies have been developed based on the main patterns that are found in the text of the comments. These patterns and an example of comment/answer are summarized in Table 1.

3. Results and discussion

The system was launched on December 31, 2014 and crawled a total of 3004 valid posts and 200,081 valid comments. The first post was dated on July 23, 2009 and the last one on December 30, 2014.

The execution of the Data retrieval module identified a total of 280 quizzes posts and 309 answer posts in the first step (quiz and answer finding process). The second step (Q&A pairing) was carried out between the sets of quizzes and answers already identified (280 and 309, respectively). This process results in the formation of 272 paired quiz-answers, but we still had 8 questions unanswered and 37 answers without a question.

A second round of this process was executed again using the remaining quizzes and answers but trying to find their match in the total number of posts extracted. This second round results in a final total of 285 paired quiz-answers with a total number of 75,215 comments attached to these quizzes. The first quiz was dated on June 11, 2011 and last one on December 29, 2014. Finally, 6 quizzes remained unanswered and 25 answers remained without a quiz and were discarded.

The execution of the fourth step (comment analysis) discards a total of 7917 comments (10.53%) from the total retrieved (75,215).

In the context of comment identification the experiment was carried out with two goals:

i. Identification of comments as valid/invalid.
ii. Identification of the answer provided in valid comments.

The aim of this approach and the analysis of the results provided were to identify the accuracy of the strategies developed for the aforementioned goals.

The first goal (identification of comments as valid/invalid) was done by means of an estimation calculated after applying a manual evaluation of precision over a sample of 260 comments chosen randomly (130 discarded and 130 not discarded and homogenously distributed among the previous 8 strategies, which were divided in 13 different options based on the

Table 1 – Summary of main patterns found in the comments.

Question
Which one of the following types of idiopathic interstitial pneumonia is associated with granulomas?
A. Acute interstitial pneumonia.
B. Cryptogenic organizing pneumonia.
C. Desquamative interstitial pneumonia.
D. Lymphoid interstitial pneumonia.

Pattern	Example
<SINGLE LETTER>	D
<SINGLE LETTER> + <PUNCTUACTION SIGNS>	"D.", "D,", "D;", "D:", ...
<SINGLE LETTER> + <EXPLANATORY TEXT>	D. A recent by the American Thoracic Society consensus statement and the European Respiratory Society recognized several distinct clinicopathological forms of idiopathic interstitial pneumonias, including idiopathic pulmonary fibrosis, ...
<TEXT OF AN OPTION/TEXT OF AN OPTION MISWRITTEN>	"Lymphocytic interstitial pneumonia", "lymphocytic interstitial neumonia", ...
<TEXT OF AN OPTION/TEXT OF AN OPTION MISWRITTEN> + <EXPLANATORY TEXT>	"Lymphoid interstitial pneumonia that occurs in some HIV cases"
<ACRONYM OF A TEXT OF AN OPTION>	"LIP", "AIP", "DIP", ...
<ACRONYM OF A TEXT OF AN OPTION> + <EXPLANATORY TEXT>	"LIP that occurs in some HIV cases"
<OTHER NO RELATED TEXT>	If you have PayPal, payza, perfect money bitcoin or solid trust so you can join this program that I will share with you if you want to earn money easily

Table 2 – Results of comment validity identification.

TP	TN	FP	FN
129	60	0	71
Precision	**Recall**	**Specificity**	**F1**
1	0.645	1	0.784

correct/incorrect identification of the strategies). The results regarding the identification of a comment as valid or invalid are summarized in Table 2.

The results reveal general good behavior. False negative (FN) values (71) are responsible for the fall in recall and F1 values. These FN have been identified as a problematic inherent to the analysis of the user's comment: several users provide large and detailed explanations regarding their answer. This means that the algorithm, using the current techniques (string processing, not advanced natural language processing) is unable to identify a valid answer in the comment, dismissing it although it contains a valid answer. However, if we take into account those comments which were correctly identified as valid answer, general results largely improve the identification of the correct answer in the comment.

The second goal was carried out by an analysis performed over the 260 chosen comments in order to estimate the efficiency of our strategies identifying the answer provided by the user's comment. The results of the analysis of this goal are shown in Table 3.

The correct answer identification has an estimated mean precision of 88%, which is a general good value. As can be seen, most of the strategies (or options inside the strategies) show very good values with a positive identification of the correct answer in 100% of the cases.

The analysis of the users' comments brings up several interesting facts. In a total of 253 posts (88.77%) the most voted answer turned out as the correct answer. Several hypotheses about the use of collective intelligence as a source of decision-making process can be extracted from these results. Following Surowiecki [28] definition about the wisdom of the crowds we could conclude that in this case, a decision based on the collective decisions of the members involved in the group could lead to a better solution than the individual opinion of its members.

A total number of 32,780 different users participated providing answers to the aforementioned quizzes. The most prolific user provided a total of 156 correct answers from a total of 194 answers provided, which means an accuracy of about 80% identifying the correct answer.

To end with, an analysis to discover trends was performed. When a Facebook user access to a post which contains a high number of comments, Facebook automatically hides most of the comments showing only the latest comments (in fact, it depends on the user configuration: Facebook can show the latest comments or only the most relevant ones). This fact makes it possible for a user to check the last answers prior to his publishing an answer to the quiz and can lead him to follow previous responses. This situation has motivated us to perform an analysis of the valid comments extracted by our system with the aim of finding "answer trends". We have used a threshold value of 5 as the number of consequent answers that should be found to consider this list of answers as a "trend".

The analysis of the quizzes and the comments reveals that, of the 285 quizzes, patterns were found in 261 (91.58%). From these 261 quizzes where trends were found, we saw that users follow trends of incorrect answers in 13 quizzes and trends of

Table 3 – Results of comment identification strategies.

Strategy		Results			
Name	Possible values	TP	FP	Precision	Mean precision
Direct Matching [1XX]	Original comment equals answer option: Correct [101]	10	0	1	1
	Original comment equals answer option: Incorrect [111]	10	0	1	
	Original comment equals answer text: Correct [102]	10	0	1	
	Original comment equals answer text: Incorrect [112]	10	0	1	
	Preprocessed comment equals answer text: Correct [103]	10	0	1	
	Preprocessed comment equals answer text: Incorrect [113]	10	0	1	
Answer found in text [2X1]	Correct answer found [201]	9	1	0.9	0.8
	Incorrect answer found [211]	7	3	0.7	
Tokenization of white spaces [3X1]	Correct answer found [301]	10	0	1	0.8
	Incorrect answer found [311]	6	4	0.6	
Levenshtein [4XX]	Correct answer by direct Levenshtein [401]	10	0	1	0.76
	Correct answer by brute force Levenshtein [402]	8	2	0.8	
	Incorrect answer by brute force Levenshtein [412]	5	5	0.5	
Final results		TP	FP	Final precision	
		115	15	0.88	

correct answers in 248. We should also remark that we have found out, in some cases, very large trends: the maximum trend size contains 182 consecutive answers in the trends of correct answers and 69 consecutive answers in the trends of incorrect answers.

4. Conclusions and future work

The presented approach allows us to discover medical knowledge in medical education context *from* the identification of the quizzes and their respective answers, with a precision rate of 88% in the identification of users' answers in the large number of posts published by NEJM in its Facebook timeline. This is a positive result because facilitates the direct extraction of quizzes which can be reused in other environments. Several conclusions could be drawn from this analysis based on several hypotheses which in fact, are really difficult to verify without asking users. Have the users followed the trend set in previous comments? Have they just copied the previous answer just because they saw a trend? Based on the size of the trends found (182 and 69 to correct and incorrect answers), the results lead us to think that this option is not totally impossible.

Another interesting aspect is that the approach proposed can be easily applied to other sources of information: the presented implementation of the system has been customized to retrieve information from NEJM Facebook timeline. However, the application to other platforms is straightforward as it only involves changing the crawling process and adapting the quiz-answer identification process (if needed). The data extracted can be used for different analysis of interest for "medical education" for instance, trying to observe patterns in order to conclude the difficulty of the tests or studying what type of questions the users have a higher or lower rate of correct answers with.

In future studies, we have two main goals: *our first goal is to apply* more complex strategies for matching users' comments and the answer options provided by the quizzes. The application of more complex natural language processing techniques such as those studied previously by several authors [29,30, p. 2] and the use of current NLP systems such as cTAKES [31] and MetaMap [32] is under consideration. The second goal is to provide a deeper (and statistically-valued) analysis of the results, emphasizing the analysis performed over the accuracy of the users involved and trying to obtain more insights from the data. We also plan to improve the storage of the analyzed data including the use of semantic technologies and biomedical ontologies to publish the data.

REFERENCES

[1] A. Rodríguez-González, M.A. Mayer, J.T. Fernández-Breis, Biomedical information through the implementation of social media environments, Biomed. Inf. 46 (December (6)) (2013) 955–956.

[2] J. Krumm, N. Davies, C. Narayanaswami, User-generated content, IEEE Pervasive Comput. 7 (October (4)) (2008) 10–11.

[3] F. Lupiáñez-Villanueva, M.A. Mayer, J. Torrent, Opportunities and challenges of Web 2.0 within the health care systems: an empirical exploration, Inf. Health Soc. Care 34 (September (3)) (2009) 117–126.

[4] G. Eysenbach, J. Powell, O. Kuss, E.-R. Sa, Empirical studies assessing the quality of health information for consumers

on the world wide web: a systematic review, JAMA 287 (May (20)) (2002) 2691–2700.
[5] Á. Leis, M.Á. Mayer, J. Torres Niño, A. Rodríguez-González, J.M. Suelves, M. Armayones, Healthy eating support groups on Facebook: content and features, Gac. Sanit. SESPAS 27 (August (4)) (2013) 355–357.
[6] D.R. George, C. Dellasega, Use of social media in graduate-level medical humanities education: two pilot studies from Penn State College of Medicine, Med. Teach. 33 (July (8)) (2011) e429–e434.
[7] D.P. Bahner, E. Adkins, N. Patel, C. Donley, R. Nagel, N.E. Kman, How we use social media to supplement a novel curriculum in medical education, Med. Teach. 34 (March (6)) (2012) 439–444.
[8] C.C. Cheston, T.E. Flickinger, M.S. Chisolm, Social media use in medical education: a systematic review, Acad. Med. 88 (June (6)) (2013) 893–901.
[9] USMLE Twitter Account. WebCite: http://www.webcitation.org/6ZftLWUXw.
[10] Crush USMLE Twitter Account. WebCite: http://www.webcitation.org/6Zfu1F3z7.
[11] Radiology Signs Facebook Page. WebCite: http://www.webcitation.org/6Zfu9eqf8.
[12] Radiopaedia.org Facebook Page. WebCite: http://www.webcitation.org/6ZfuF9L87.
[13] New England Journal of Medicine, Image Challenge Facebook App. WebCite: http://www.webcitation.org/6ZfuL6mIm.
[14] New England Journal of Medicine, Interactive Medical Case. WebCite: http://www.webcitation.org/6ZfuSS5Eu.
[15] Medical Quiz Facebook Page. WebCite: http://www.webcitation.org/6ZfuY0X5Q.
[16] Medical Quizzes Facebook Page. WebCite: http://www.webcitation.org/6ZfueeruK.
[17] S.-M. Kim, E. Hovy, Extracting Opinions, opinion holders, and topics expressed in online news media text, in: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Stroudsburg, PA, USA, 2006, pp. 1–8.
[18] M.A. Shrivastava, B. Pant, Opinion extraction and classification of real time Facebook Status, Glob. J. Comput. Sci. Technol. 12 (8) (2012 Apr.).
[19] J. Kim, D. Choi, B. Ko, E. Lee, P. Kim, Extracting user interests on Facebook, Int. J. Distrib. Sens. Netw. 2014 (June) (2014) e146967.
[20] B. Rieder, Studying Facebook via data extraction: the Netvizz application, in: Proceedings of the 5th Annual ACM Web Science Conference, New York, NY, USA, 2013, pp. 346–355.
[21] P. Gundecha, H. Liu, Mining social media: a brief introduction, in: New Directions in Informatics, Optimization, Logistics, and Production, INFORMS, 2012, pp. 1–17.
[22] C. Aliprandi, A.E. De Luca, G. Di Pietro, M. Raffaelli, D. Gazze, M.N. La Polla, A. Marchetti, M. Tesconi, CAPER: crawling and analysing Facebook for intelligence purposes, in: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2014, pp. 665–669.
[23] S.A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Crawling Facebook for social network analysis purposes, ArXiv11056307 Phys. (2011) 1.
[24] A. Rodríguez González, M.A. Mayer-Pujadas, E. Menasalvas-Ruiz, NEJM FB Source Code. WebCite: http://www.webcitation.org/6Zfuo6T3r.
[25] Facebook, Facebook Developers webpage. WebCite: http://www.webcitation.org/6Zfutl8VG.
[26] Facebook, Graph API – Facebook developers. WebCite: http://www.webcitation.org/6Zfuy6O5L.
[27] Apache Foundation, Properties file. WebCite: http://www.webcitation.org/6Zfv37Mcn.
[28] J. Surowiecki, The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations, Doubleday, 2004.
[29] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inf. Assoc. 1 (2) (1994) 161–174.
[30] S. Goryachev, M. Sordo, Q.T. Zeng, A suite of natural language processing tools developed for the I2B2 project, AMIA Annu. Symp. Proc. 2006 (2006) 931.
[31] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inf. Assoc. JAMIA 17 (October (5)) (2010) 507–513.
[32] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proc. AMIA Symp., 2001, pp. 17–21.