Medical Information Extraction Model for User-generated Content

Fahad Kamal Alsheref

Information System Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt

Corresponding author: Fahad Kamal Alsheref, PhD. Information System Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt. Address: Cairo, Egypt. E-mail address: Drfahad@fcis.bsu.edu. ORCID ID: <https://www.orcid.org/0000-0001-5894-9583>,

doi: 10.5455/aim.2019.27.192-198

ACTA INFORM MED. 2019 SEP 27(3): 192-198

Received: Jul 10, 2019 • Accepted: Aug 10, 2019

© 2019 Fahad Kamal Alsheref

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Introduction: The number of social network users is on the rise, and the size of the user-generated contents is increasing as well. Analyzing the generated contents can lead to the attainment of a vast amount of information, such as users' feelings on specific products or events, or personal information about life events. **Aim:** The aim of this paper is to describe an model for detecting medical information present in generated contents, such as posts or comments. **Results:** The proposed model is based on the Unified Medical Language System (UMLS) and is tested on a dataset collected from Twitter and Facebook. The extracted information can be used to aid in the early detection of diseases or to supply commercial benefits to medical companies. Experimental results demonstrate that the proposed model achieves 94.6% accuracy and 87% precision. **Conclusion:** In this study, we attempted to extract clinical information present in UGC. Using the proposed model should involve a reliable dataset that contains most clinical expressions; the UMLS was a suitable dataset for our model.

Keywords: Text similarity, Electronic health record, Facebook, Social network.

1. INTRODUCTION

Social network platforms have become a vital part of people's lives; through them, people express their opinions and describe daily events. Web applications based on Web 2.0 technology encourage user participation through the contents generated by the users. For instance, individuals may create a post or comment about their life, events, stories, or medical conditions (1). Such content is called user-generated content (UGC), which gradually increases over time. The UGC contains valuable information that could be used in several applications such as question answering, blog or review mining, and information extraction about a specific domain. Social networks contain considerable amount of UGC, such as posts, reviews, and comments. UGC is a publicly available media content that are produced by end-users, which is made without standard content and format; moreover, validating the contents is not possible, which is an important topic in research for measuring the UGC credibility (1).

UGC may be in a structured and unstructured format; structured format such as author and publication date is applied in a template, and unstructured format is a free text without any template or structure; most of UGCs are in unstructured format, in which detecting and identifying the information is difficult (2).

Detecting and analyzing medical information in individuals' posts can be used as a warning to physicians or an alarm regarding infections in the given regions. It can also be used to evaluate the effects of certain drugs (2). Several techniques used to detect the information in UGC are natural language processing (NLP) techniques, text mining, and data mining (3).

The main principle of our study involves using existing techniques for detecting medical information from individual's posts; this detection is based on the Unified Medical Language System (UMLS) repository, which is explained in the following sections.

2. AIM

The aim of this paper is to describe an model for detecting medical information present in generated contents, such as posts or comments.

3. METHODS

There are several techniques used in our proposed model, including text mining, the vector space model, and UMLS. These are discussed below.

Text mining

Text mining is a branch of data mining that involves searching for hidden information in a text corpus; in other words, it is the process of extracting valuable information from text (4). This information is typically derived through several steps as follows (5):

1. Text preprocessing;
2. Part-of-speech tagging;
3. Statement segmentation;
4. Noun phrase extraction.

Text preprocessing is referred to as tokenization and consists of the following steps (6):

- Discarding unwanted elements, such as brackets and tags.
- Processing word boundaries (whitespace and punctuation).

Stemming, or extracting words' original forms. For example, the English word *look* can be inflected with morphological suffixes to produce *looks*, *looking*, and *looked*. These words share the same stem: *look*. Stemming is a complex process, as there can be many exceptions (e.g., department vs. depart, be vs. were). The most commonly used stemmer is the Porter stemmer (7).

Removing stop words: the most frequently used words often carry little meaning.

Capitalizing and case folding. It is often convenient to convert all characters to lowercase.

Part-of-speech tagging involves software that reads a text in a given language and assigns parts of speech to each word, such as nouns, verbs, and adjectives (8).

Statement segmentation serves to divide the text into several statements (9).

Noun phrase extraction is responsible for extracting noun phrases; complex noun phrases are then decomposed into simpler noun phrases (9).

Vector space model

A vector space model is an algebraic model for representing text documents as vectors of identifiers, such as

index terms. It is used in information filtering, information retrieval, indexing, and relevance ranking (10). A common use of this algorithm is classification, which is achieved by measuring the similarity between two texts or documents. Each document is represented as a vector, and the cosine of the angle between them is calculated. The closer the value is to 1, the higher the similarity between the documents (11).

Clinical data is classified into several classes, including etiology, complaints, procedures, diagnosis, prognosis, treatment, and prevention. Each class is defined by important keywords, and the similarity between a user's text and the clinical class is measured through shared keywords (12). A clinical phrase is correlated to each class by the following equation:

$$\text{Norm}(P) = \sqrt{\sum W(j)^2} \quad (12),$$

where $W(j)$ is the weight of the word phrase in the defined class.

$$\text{Cosine}(P1, P2) =$$

$$\frac{\sum (wp1(j) * wp2(j))}{(\text{Norm}(P1) * \text{Norm}(P2))} \quad (12),$$

where wpi is the weight of a word phrase for class i . The cosine similarity between the phrase and class ranges from 0 to 1, and the angle between two-term frequency vectors cannot be greater than 90°. Thus, the closer the cosine value is to 1, the more similar the clinical phrase is to the class (13).

UMLS

The UMLS is a collection of files and software that consists of nearly all health and biomedical vocabularies and standards (14). Thus, it is an extensive collection of many controlled vocabularies in the biomedical sciences and provides a mapping structure for these vocabularies, thereby enabling translation across various terminology systems (15). The UMLS consists of the following components (16):

Metathesaurus: the primary database of the UMLS that includes a collection of concepts and terms from various controlled vocabularies, and their relationships.

Semantic Network: a set of categories and relationships that are used to classify and relate entries in the Metathesaurus.

SPECIALIST Lexicon: a database of lexicographic information to be used in NLP.

Each medical phrase is registered in the UMLS repository with a description and relation to other phrases. Extracted phrases from UGCs are examined in the UMLS to determine their meanings and relations to other phrases to extract users' medical information (17).

Classification of medical information

In the UMLS, medical expressions are classified into three main areas:

Examination: the process of investigating the body of a patient for signs of disease by medical professionals (18).

Diagnosis: the process of determining the disease or condition that can explain a person's symptoms and signs (19).

Procedure: a collection of actions intended to achieve a

Figure 1. Vector space model (11)

No	UMLS Code	Examination	Diagnosis	Procedure
1	Clinical Drug	0	0	1
2	Finding	1	0	0
3	Laboratory	0	0	1
4	Test Result	0	0	1
5	Sign or Symptom	1	0	0
6	Virus	0	1	0
7	Disease	0	1	0
8	Syndrome	0	1	0
9	Vitamin	0	0	1
10	Organism Function	1	1	0
11	Neoplastic	0	0	1
12	Process	0	0	1
13	Mental Dysfunction	0	1	0
14	Behavioral Dysfunction	0	1	0
15	Mental Process	0	0	1
16	Hormone	0	0	1

Table 1. UMLS codes and its medical classes

result in the delivery of healthcare (20).

Each area is treated as a class and has its primary expressions. The first step in the classification process is to build a collective set of features, typically called a dictionary. The dictionary of words covers the majority of possible medical expressions and their suggested classes. Table 1 illustrates an example of the classification dictionary.

We implemented the text preprocessing and classification process, which was presented in (21). The second step is to measure the similarity between the extracted expressions and the predefined classes and to classify each expression into the most appropriate class. Cosine similarity is the algorithm used in this study to calculate similarity.

4. RESULTS

There are several studies in the field of medical information extraction, and they use a variety of natural languages. Chen et al.(22) proposed a model to extract clinically useful information from Chinese electronic medical records. In particular, they developed an NLP-based algorithm for extracting clinical information regarding patients with hepatocellular carcinoma (HCC) from these records. Their model focused on clinical information present in operation notes as well as radiology and pathology reports. Collected from 92 HCC patients, this dataset was divided into a training set of 60 patients and a test set of 32 patients to evaluate the model. Rule-based and hybrid methods were used for extracting information, and the dataset set was manually annotated to measure the performance of the model. The performance was measured by calculating the precision, recall, and F-score, all of which had a score of $\geq 80\%$ (22). Thus, the model proved to be successful, but with limitations: only specific types of documents relating to specific diseases were used, and this model focuses only on the Chinese language. It would be helpful to generalize this model to apply to a broader range of clinical documents and other natural languages.

Bushinak et al.(21) presented a model for extracting medical information from free text. The free text may be a patient's report or a prescription. They tried to convert the unstructured medical information to a structured format by identifying conditions such as disease symptoms. They used Text mining and NLP techniques for identifying medical information (21).

Tang et al.(23) tried to identify and track topics discussed on a cancer institution's Facebook page and extract useful information about emotional support to patients and family members in the free text UGC. They classified the extracted information into greetings and comments about the cancer institution, blessings, time, treatment, expressions of optimism, tumor, father figure, and other family members and friends, and the other comments were unclassified. This research confirms the importance of the UGC, and it is used as a source of structured information after applying information extraction process (23).

Xinying Song et al. (24) proposed an enhanced model of mining data records (MDR) in Web pages, the original MDR is based on two key observations about the layout of data records in Web pages and uses a string-matching algorithm (24). He adopted the domain constraints to enhance the string similarity, but their work focused on the web pages in general. Working on social networks UGC is different because in social media such as Facebook, people express their opinions and feelings that are related to their life. Conversely, MDR focuses on extracting information from the web pages to put a similar text together in structured or semi-structured formats (24).

Sean D. Young et al. used social media as an early indicator of syphilis (25). They utilized the increasing number of social media users and the inexpensiveness of collecting data from social media. The goal of their proposed model was to work as a cost-effective surveillance strategy of syphilis disease. The data were collected from Twitter, and they were filtered to include only sex-related tweets from the United States; words that contain sexual meaning such as "sex" and "fuck" were selected to be associated with sexual risk-related attitudes and behaviors. However, their list of words is limited and does not guarantee the existence of syphilis as they do not include other medical expressions that could be a good predictor of this disease; thus, applying our method may enhance their proposed algorithm (25).

Viani et al. (26) attempted to extract information from Italian medical reports using an ontology-driven approach; their goal was to identify events and their attributes from medical reports written in Italian. They built a corpus that included 5,432 non-annotated medical reports about patients with rare arrhythmias. For extracting clinical information, they built a domain-specific ontology that included events and attributes to be extracted with predefined regular expressions. The proposed model performance was evaluated on an independent test set and achieved an accuracy of 90% for most clinical cases. This model succeeded in extracting clinical information from Italian records; however, it was limited to a specific domain and language (26).

Chiaramello et al.(27) studied information extraction from Italian medical documents using „off-the-shelf“ information extraction algorithms. They conducted three experiments that demonstrated that the Italian UMLS Metathesaurus sources covered 91% of medical expressions in the Italian clinical notes. These results reinforce the importance of the UMLS as a verified source of clinical expressions (27).

In our study, we focused on UGCs, especially content obtained from social media platforms, such as Facebook and Twitter. The UGC can be used as an early alarm for infections in a specific region, or as a marketing tool for doctors and pharmaceutical companies (27).

4.1. Proposed model

The proposed model consists of multiple steps, as illustrated in Figure 2 and described below:

- The user creates UGC, which can be a post or a tweet.
- The UGC is input to a text mining process that is responsible for extracting noun phrases after applying text preprocessing, part-of-speech tagging, and statement segmentation.
- The extracted phrases are inputted to a process for searching the UMLS repository.
- The presence of medical information in the posts is determined, and the posts are classified into one of the classes mentioned in Section 2.4.

Figure 2. Proposed model with its steps

4.2. Evaluation

The dataset was built by selecting a list of UGC Face-

book posts and Twitter tweets; it contained 500 UGCs collected from 500 users (250 Facebook users and 250 twitter users). We used a version of the Twitter API based on python and c#. which pulled queries from Twitter's public timeline, but for Facebook, the posts collected manually because of Facebook policy that prevents the API usage. Table 2 presents a snapshot of the collected dataset.

UGC ID	UGC	Source	User ID
1	My muscles are sore.	Facebook	1
2	My nose is stuffy.	Facebook	2
3	My silence/smile is just another word for my pain.	Twitter	3
4	My stomach hurts.	Facebook	4
5	Never underestimate the power of denial, the heights of assumption or the depths of pain.	Twitter	5
6	Diabetes is a part of my life, but that does not mean I have to love it	Facebook	6
7	Our health always seems much more valuable after we lose it.	Facebook	7
8	Pain is the only thing that's telling me I am still alive.	Twitter	8
9	People cry, not because they are weak. It is because they have been strong for too long.	Facebook	9
10	Yes, in my diabetes lifetime, I have stuck a needle in my fingertips	Twitter	10

Table 2. Dataset snapshot

There are eight volunteer physicians in different clinical specializations that participated in the research, they are listed in Table 3. Each UGC was manually annotated and classified by the participated physicians. Each class was defined through a vector of expressions, as demonstrated in Table 1. The UGC was entered into the proposed model, as illustrated in Figure 2, and the extracted medical expressions were classified into predefined classes.

No	Specialization	Degree
1	Specialty Ophthalmology	Ph.D.
2	General Surgery	Ph.D.
3	Specialty Oncology Surgery	Ph.D.
4	Audiology specialization	M.Sc.
5	Cardiology specialization	M.Sc.
6	Specialty Pediatrics	M.Sc.
7	Specialty Orthopedic Surgery	M.Sc.
8	Dermatology and Genetics	M.Sc.

Table 3. Physicians who participate in manual annotations

5. DISCUSSION

The following example demonstrates the complete journey of a created post through the proposed model up to the completion of the classification process, according to the cosine similarity and vector space model. The following post is considered:

Diabetes is a part of my life, but that does not mean I have to love it

The text mining package that used in our proposed

model is Natural Language Toolkit (28), it is a leading platform for building Python programs to work with human language data and could be integrated with Microsoft platforms. This post enters the text mining process, upon which the following subprocesses are applied:

- Text preprocessing: all brackets, unwanted features, and word boundaries are removed.
- Part-of-speech tagging: parts of speech are assigned to each word.
- Statement segmentation: examination text is split into multiple statements.

The output of the process is provided in Table 4.

Word	Lemma	Tag
diabetes	diabetes	Noun, singular or mass
part	part	Noun, singular or mass
life	life	Noun, singular or mass

Table 4. Extracted noun phrases

Extracted noun phrases continue to the following step and are converted to UMLS codes. The UMLS API is responsible for determining the associated class for each noun phrase; the output of this process is presented in Table 5.

Noun Phrase	UMLS code
diabetes	Disease
part	Noun, singular or mass
life	Noun, singular or mass

Table 5. Example of Noun phrases and their UMLS code

The vector space model is then used to measure similarity to determine the most appropriate class corresponding to the extracted clinical information. The cosine values for the three classes are as follows:

$\text{Cos (Examination)} = 0$

$\text{Cos (Diagnose)} = 0.408$

$\text{Cos (Procedure)} = 0$

The Diagnose class has the most significant value; it is thus the winning class. Table 6 presents sample terms that are manually and automatically annotated.

After extracting clinical information from the UGC, it is classified into predefined classes, which are defined manually and by the proposed model. Table 6 presents a subset of model classifications and manual classifications.

To measure the performance of the proposed model, we calculated the precision, recall, and F-score with the following equations (29):

Precision: $(P) = TP / (TP + FP)$

Recall: $(R) = TP / (TP + FN)$

F-score = $2PR / (P + R)$

Here, TP denotes true positive, FP denotes false positive, and FN denotes false negative. For each UGC entry, we identified the TP, FP, and FN to calculate the precision, recall, and F-score. Tables 8 and 9 summarize the results of applying the model to 500 UGCs and presents the values that measure model performance.

The accuracy of classification = $(\text{No. of true classifications} / \text{total number of UGCs}) = (472/500) * 100 = 94.6\%$.

ID	UGC	Extracted terms with manual annotations	Extracted terms with proposed model annotations
1	My muscles are sore.	muscle (human part) Sore (Finding)	muscle (human part) Sore (Finding)
2	My nose is stuffy.	nose (human part) stuffy (Finding)	nose (human part) stuffy (Finding)
3	My silence/smile is just another word for my pain.	None	Pain (Finding)
4	My stomach hurts.	stomach (human part) hurt (Finding)	stomach (human part) hurt (Finding)
5	Never underestimate the power of denial, the heights of assumption or the depths of pain.	None	Pain (Finding)
6	Diabetes is a part of my life, but that does not mean I have to love it.	diabetes (disease)	diabetes (disease)
7	Our health always seems much more valuable after we lose it.	None	None
8	Pain is the only thing that's telling me I'm still alive.	None	Pain (Finding)
9	People cry, not because they're weak. It's because they've been strong for too long.	None	None
10	Yes, in my diabetes lifetime, I have stuck a needle in my fingertips.	diabetes (disease)	diabetes (disease)

Table 6. UGC Manual annotation and model annotation

ID	Post	Manual Classification	Model Classification	Result
1	My muscles are sore.	Examination	Examination	True
2	My nose is stuffy.	Examination	Examination	True
3	My silence/smile is just another word for my pain.	Examination	Examination	True
4	My stomach hurts.	Examination	Examination	True
5	Never underestimate the power of denial, the heights of assumption or the depths of pain.	No Class	Examination	False
6	Diabetes is a part of my life, but that does not mean I have to love it.	Diagnose	Diagnose	True
7	Our health always seems much more valuable after we lose it.	None	None	False
8	Pain is the only thing that's telling me I'm still alive.	None	Examination	False
9	People cry, not because they're weak. It's because they've been strong for too long.	None	None	False
10	Yes, in my diabetes lifetime, I have stuck a needle in my fingertips.	Diagnose	Diagnose	True

Table 7. Post manual classification and model classification

6. CONCLUSION

UGC ID	TP	FP	FN	Precision	Recall	F-score
1	1	0	0	100.00	100.00	100.00
2	1	0	0	100.00	100.00	100.00
3	1	0	0	100.00	100.00	100.00
4	1	0	0	100.00	100.00	100.00
5	3	0	0	100.00	100.00	100.00
6	1	0	0	100.00	100.00	100.00
7	2	1	1	66.67	66.67	66.67
8	1	1	0	50.00	100.00	66.67
9	1	1	0	50.00	100.00	66.67
10	1	0	0	100.00	100.00	100.00

Table 8. Performance measure values

Precision	Recall	F-score
87.00	89.08	85.32

Table 9. Average precision, recall, and F-score

Social media generates billions of data points that can be used as a vital source of information. In this study, we attempted to extract clinical information present in UGC. Using the proposed model should involve a reliable dataset that contains most clinical expressions; the UMLS was a suitable dataset for our model. After applying the proposed model, we measured its performance and observed 94.2% accuracy, 87% precision, 89% recall, and an 85.32% F-score. These results demonstrate the success of our proposed model in extracting and classifying the medical information.

- **Clinical Relevance Statement:** This research focuses on finding the hidden clinical information that exists in UGCs. This information could be useful in early detection of diseases or the clinical marketing process.
- **Protection of Human and Animal Subjects:** The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.
- **Author's Contribution:** F.K.A. gave substantial contribution to the conception or design of the work and in the acquisition, analysis and interpretation of data for the work. Author had main role in drafting the work and revising it critically for important intellectual content and gave final approval of the version to be published and they agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
- **Conflict of Interest:** There are no conflicts of interest.
- **Financial support and sponsorship:** Nil.

REFERENCES

1. Couldry N. Media, society, world: Social theory and digital media practice. Polity. 2012.
2. Saha L. Irritable bowel syndrome: pathogenesis, diagnosis, treatment, and evidence based medicine. World Journal of Gastroenterology: WJG; 2014; 20(22): 6759.
3. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems. 2013; 28(2): 15-21.
4. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier, 2011.
5. Santos CD, Zadrozny B. Learning character. Level representations for part of speech tagging. In Proceedings of the 31st International Conference on Machine Learning (ICML.14).2014: 1818-1826.
6. Bushinak H, AbdelGaber S, AlSharif FK. Recognizing the electronic medical record data from unstructured medical data using visual text mining techniques. International Journal of Computer Science and Information Security. 2011; 9(6): 25.
7. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill Manning CG. KEA: Practical Automated Keyphrase Extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. 2005: 129-152.
8. Voutilainen A. Part of speech tagging. The Oxford handbook of computational linguistics. 2003: 219-232.
9. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates. 2001: 71.
10. Yan TW, Garcia Molina H. Index structures for information filtering under the vector space model. In Data Engineering. Proceedings of 10th International Conference. 1994: 337-347.
11. Dhillon IS, Modha DS. Concept decompositions for large sparse text data using clustering. Machine learning. 2001; 42(1-2), 143-175.
12. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo Jr, JL, Roccella EJ. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. hypertension. 2003; 42(6): 1206-1252.
13. Widdows D, Cohen T. Reasoning with vectors: A continuous model for fast robust inference. Logic Journal of the IGPL. 2014; 23(2): 141-173.
14. Butte AJ, Kohane IS. Creation and implications of a phenotype genome network. Nature biotechnology. 2006; 24(1): 55.
15. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association. 1993; 81(2): 217.
16. Aronson AR. Metamap: Mapping text to the umls metathesaurus. Bethesda, MD: NLM, NIH, DHHS. 2006: 1-26.
17. Chen H, Lally AM, Zhu B, Chau M. HelpfulMed: intelligent searching for medical information over the internet. Journal of the American Society for Information Science and Technology. 2003; 54(7): 683-694.
18. Lupton D. Medicine as culture: illness, disease and the body. Sage. 2012.
19. O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health services

- research. 2005; 40(5p2). 1620-1639.
20. Nieva VF, Sorra J. Safety culture assessment: a tool for improving patient safety in healthcare organizations. *BMJ Quality & Safety*. 2003; 12(suppl 2): ii17-ii23.
21. Bushinak H, Abdel Gaber S, AlSharif FK. Recognizing the electronic medical record data from unstructured medical data using visual text mining techniques. *International Journal of Computer Science and Information Security*. 2011; 9(6): 25.
22. Chen L, Song, L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. *International Journal of Medical Informatics*. 2019; 124: 6-12.
23. Tang C, Zhou L, Plasek J, Rozenblum R, Bates D. Comment topic evolution on a cancer institution's Facebook page. *Applied clinical informatics*. 2017; 8(03): 854-865.
24. Song X, Liu J, Cao Y, Lin CY, Hon HW. Automatic extraction of web data records containing user-generated content. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010: 39-48).
25. Young SD, Mercer N, Weiss RE, Torrone EA, Aral SO. Using social media as a tool to predict syphilis. *Preventive medicine*. 2018; 109: 58-61.
26. Viani N, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R, Sacchi L. Information extraction from Italian medical reports: An ontology driven approach. *International journal of medical informatics*. 2018; 111: 140-148.
27. Chiaramello E, Pincioli F, Bonalumi A, Caroli A, Tognola G. Use of "off the shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of biomedical informatics*. 2016; 63: 22-32.
28. Natural Language Toolkit (<https://www.nltk.org/>)
29. Powers DM. Evaluation: from precision, recall and F.measure to ROC, informedness, markedness and correlation. 2011.

EFMI

EUROPEAN FEDERATION
for MEDICAL INFORMATICS

[Home](#)
[About](#)
[Working Groups](#)
[Institutional Members](#)
[Events](#)
[Conferences](#)
[Publications](#)
[Contact](#)
[More](#)

Upcoming Events

No events

Login

☐ Remember Me

[Forgot your password?](#)

MEDINFO 2019
 HEALTH AND WELLBEING E-NETWORKS FOR ALL www.medinfo-lyon.org
 CONVENTION CENTRE 25-30 AUGUST 2019, LYON FRANCE

AIM (the French Medical Informatics Association) is pleased to host the 17th world congress of medical and health informatics in Lyon (France). The conference is organized on August 25th to 30th 2019.
[More information at http://www.medinfo-lyon.org/en/](http://www.medinfo-lyon.org/en/)

Save the date: GNUHealthCon 2019

Created: 19 June 2019

Join us at GNU Health Con 2019 in Liege, Belgium, 13-15 December

GNU Health Con is an annual conference that brings together enthusiasts and developers of the Free/Libre Health & Hospital Information System.

This year GNUHealthCon joins forces with the Orthanc community, making an even stronger Libre Software community in the healthcare arena. During the 3-day conference, you'll have the opportunity to attend the different thematic sessions and free workshops, share experiences, implementation cases and, of course, network with GNU Health community members from around the world! Participation is free of charge but a donation for the conference organizing team is very welcome. Find more information here: <http://gnuhealthcon.org/2019-liege/>