1-1-2016

Mining the Web and Literature to Discover New Knowledge about Diabetes

Farhi Marir
Zayed University

Huwida Said
Zayed University

Feras Al-Obeidat
Zayed University

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 1256 – 1261

Procedia Computer Science

The 3rd International Workshop on Machine Learning and Data Mining for Sensor Networks (MLDM-SN)

Mining the Web and Literature to Discover New Knowledge about Diabetes

Farhi Marir*[a], Huwida Said[a], and Feras Al-Obeidat[a]

[a]*College of Technological Innovation, Zayed University, UAE*

Abstract

Social Networks are powerful social media for sharing information about various issues and can be used to raise awareness and collect pointers about associated risk factors and preventive measures in chronical disease like diabetes. Since the olden times, knowledge in medicine was established through recording and analysing human experiences. This paper presents the results of text mining techniques of more than five hundred thousands of texts retrieved from social networks, blogs, forums, and also research papers from MEDLINE database to discovering new knowledge related to diabetes disease covering symptoms and treatments. The text mining approach consists of two tasks, descriptive and predictive. The descriptive task was to identify explicit references to the diabetes diseases diagnosis and treatments, whereas the predictive task focused on the prediction of the diabetes disease status when the evidence was not explicitly asserted. The findings are then compared to the standard diabetes diagnosis and treatments and only those which are not listed in the standards are retained as hypothesis for further validation by clinicians and medical researchers in the domain of diabetic disease.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Text Mining; Research papers; Diabetes; Social Networks,

1. Introduction

Early signs and symptoms of diseases such as diabetes in most cases are easily dis-regarded or brushed aside as something of a minor concern. It is the same case with other known diseases that have predominantly led to many

* Corresponding author. Tel.: +971-4402 1598; fax: +971-4402 1017.
E-mail address: Farhi.Marir@zu.ac.ae

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.261

losses of lives. It has been reported that around two-thirds of people with diabetes might be dying prematurely from complications that could have been prevented or delayed. Diabetes UK has claimed that thousands of patients were being struck down by heart attacks, strokes, kidney disease because they were diagnosed with the underlying condition too late [1]. In addition, conditions such as the type 1 diabetes is known to be five times more common than meningitis, but the diagnosis is often delayed (UK Children with Diabetes Advocacy Group). Therefore, knowledge or awareness of the early symptoms could not only help in keeping the disease under control, but it may also help in preventing any further damage. Sometimes people with a certain condition such as Diabetes type 2 are asymptomatic, however it is possible that the patient has experienced symptoms but was either mild or developed gradually, and thereby went unnoticed [2]. Internet has become a popular platform where vast amounts of information and data have been gathered every day. People use this platform to confide every facet of their life experiences in blogs, websites, forums etc. This sea of data can only be expected to increase with the popularity of the Internet, and has significant potential economic and societal value. Text mining techniques are complex and innovative methods for analysing data and they could be used to exploit this potential of Internet information. The focus of this research is on mining social network stories of people telling their experiences in being diagnosed with diabetes and also MEDLINE® database which contains journal citations and abstracts for biomedical literature from around the world. The goal is to understand and test whether such type of free-form text could provide any useful links or patterns in discovering the conditions earlier, or could provide some insight on new pattern that were not considered or recognized previously.

This paper is composed of four sections. Section 2 is devoted to presenting previous work on text mining and its application to the domain of health. Section 3 is devoted to presenting the text mining approach we use to process and analyse social network and MEDLINE research papers texts to discover knowledge in diabetes disease and the final sections are devoted to conclusions and results.

2. Previous Research Work

The concept of applying training techniques to analyse a given text for identifying patterns in the field of medical science has been attempted before, although the majority of it has been done on clinical records, biomedical texts, and documents. Like for instance the developed information retrieval system that offers semantic searches on diabetes disease across MEDLINE abstracts using 'semantic metadata' derived from text mining[3] and the research work reported in[4] on unseen patients discharge summaries. It is assumed that the data used in these research work [3,4] were more organized in their structure than the one being attempted for in this research. Clinical texts, and documents would at the very least have some level of coherence than the one's obtained from online blogs and forums which no doubt consists of noise, and poor language structure. In addition, several researchers have studied the use of appropriate text mining algorithms and concepts to support ontology engineering [5, 6] as well as its applications in diverse fields[7,8]. One of the notable works that is quite similar is the use of developed text mining techniques covering self-organizing maps (SOM), and support vector machines (SVM). This is to develop a prospective system for automating the extraction of relationships between cancer diseases and potential factors from clinical records [8], which differs from the majority of the other works where co-occurrence analysis has been the preferred choice of approach, for example in the research conducted to evaluate knowledge discovery of disease-disease relationships for rheumatic diseases [9]. Research conducted previously in text mining of clinical records were more concentrated towards text classification or Named Entity recognition, such as the work where the focus is on the extraction of textual attributes from raw text documents, and then mapping them to a structured knowledge sources, all the while relying on predefined ontology [10]. The results of this work showed that this system did perform well in terms of precision and recall, but it fell behind in extending this fair performance towards ontology matching which showed lower values. One common reason for this failure may lie in the proper extraction of candidates, since it is a complex process, but there is also the question of whether this kind of system would be able to handle unconventional writing style, or complicated sentence structures which may be an issue during the course of this research. This grey area has been addressed in another work, which employs the approach of using a verb-centric algorithm unlike previously where a co-occurrence pattern was used, to extract knowledge from biomedical text [11]. Here, the main verbs, i.e. the action verbs are identified and extracted with which two involved entities of a relationship are extracted. This approach has been claimed to work better than the previous ones with a significantly

higher F-score, and is apparently quite effective in dealing with complex sentence structures such as the clauses and conjunctive sentences. The applied concept also seem to address the limitations that other methods are known to exhibit, such as a low recall issue, severely limited ability to uncover relevant relations, expensive to construct, etc., by proposing to extract the main verbs from a sentence where a main verb is the most important verb in a sentence. Any further problems with unconventional writing style could be minimized and corrected by improving the algorithm to identify and extract prepositions. This approach has further been supported in building an opinion mining system with higher precision rate and recall value [12]. It has been disputed that applying this method of approach may not work effectively in cases where people use mixed languages which is, however not relevant to this research work.

3. The Diabetes Mining Experiment

Text mining is the process of turning text into data that can be analysed. With the advent of digital text, it allows the machines to perform text mining faster and more consistently than being done manually [13]. Text mining has its roots in computational linguistics and Information Retrieval [14]. Text mining can also be referred to as the discovery of new knowledge from a large body of text using computational algorithms to extract semantic logic [15]. In this paper, we used text mining technique to mine social networks and research papers to discover new diagnosis and treatment of Diabetes patients. The decision to use these data sources is further motivated by the fact that any existing medical knowledge so far, are based or established from human experiences. The kind of lifestyle that a person has led up to the day of his diagnosis could provide insights into any factors that were unrecognized or ignored previously. This is consolidated by the Swanson [16] who used basic text mining techniques by analysing different texts to suggest a provocative connection between dietary fish oil and Raynaud's disease, a circulatory disorder, which three years later was validated through clinical trials. We used keyword retrieval approach to retrieve around five hundred thousand stories and research papers from social networks and MEDLINE database combining different keywords related to diagnosis and treatment of diabetes. Each of the stories and research paper is stored as individual text files. From these text files we used our text processing tools developed in our big data analytics research lab known as *Kahina (Magic in Arabic)* text analytics tools to develop two indexed corpuses; one for social network stories and one for research papers. The text mining method we used is adapted from [4] and it consists of breaking down the text files into millions of sentences on which two predictive approaches are undertaken: descriptive and intuitive. The descriptive task was to identify explicit references to the diagnosis and treatment of diabetes diseases in the narrative texts files residing in the corpuses and the predictive task which focuses on inferring the diagnosis/symptom and treatment of diabetes when the evidence is not explicitly asserted.

3. 1 Descriptive Prediction

The main objective of this module was to retrieve sentences that mention explicitly "Diabetes" term along the associated terms like "diagnosis" or "treatment" and their respective associated synonyms and clinical terms. We lexically profiled diabetes disease by collecting (i) its name and synonyms from public resources including the UMLS4, (ii) diabetes sub-classes (e.g., *diabetes type I & II*) and their synonyms, (iii) disease super classes (e.g., obesity) and their synonyms, and (iv) clinical terms closely related to the diabetes disease, imported from public medical resources. Initially, the sentences that contained any term from the lexical profile were retrieved, and, in the subsequent steps, the evidence was checked against a look up table of known diabetes diagnosis and if it is found then it is filtered out otherwise it is retained with an assumption that it is a discovery. The sentence-based predictions were then combined at the corpus level. The two processing steps in this module are described briefly below.

Step #1: Term Association. To implement this approach we used the lexical module of *Kahina* text analytic tool that gives the frequency of occurrence of a given term with all other terms in the corpus. Then using the text processing module of *Kahina* text analytic, we select the highest frequency of occurrence to retrieve all sentences in each file of the corpus where the two terms occur together. In Fig. 1 below we extracted all sentences that contain "diabetes" term along "diagnosis" and similarly sentences that contain the "diabetes" term along "treatment". This process is also repeated for individual e.g. symptom and thirst and combination of synonyms and clinical terms

related to diabetes, diagnosis, and treatment terms. To cater for terminological variation, terms that characterize Diabetes disease were matched against the text approximately, taking into account morphological variants, ignoring word order.

file2928247 Urine and blood tests can be used to confirm a *diagnosis* of diabetes based on the amount of glucose in the urine and blood.

file2928247 A plasma glucose level of 2000 mg/L or higher at two hours after drinking the syrup and at one other point during the two-hour test period confirms the *diagnosis* of diabetes [7].

file2928248 The earlier *diagnosis* of diabetes , risk of the complications can be dodged.

file1603685 The case exhibiting more symptoms and signs of deficiency of the kidney-yin Diabetes is marked as feverish sensation in the palms and soles, restlessness, red tongue with little fur, deep, thread and rapid pulse; while the case presenting more symptoms and *signs* of deficiency of the kidney-yang is manifested as light colour urine, aversion to cold, impotence, pale tongue with whitish fur, and deep, thread and weak pulse.

file1603666 Meeting with a dietician is highly recommended for this type of *diabetic* natural treatment .

file1603666 Hydrotherapy -Hydrotherapy has been proven quite effective as a *diabetic* natural treatment .

file1603666 This *diabetic* natural treatment is found to be effective because Alpha Lipoic Acid has antioxidant properties which reduce those bodily abnormalities which cause a reduction of endoneural blood flow and oxygen tension.

Fig 1. Example of Sentences with co-occurrence of terms associated to Diabetes diagnosis.

Step #2: Filtering Discovery Sentences. Once the Term association sentences are collected as shown in Fig. 1 above, two processes of filtering out similar sentences are undertaken. *Kahina* text analytic tool implements basic sentence similarity algorithms by partitioning sentence into a list of tokens, then undertake Medical POS tagging, and perform stemming word algorithms and finally compute the similarity of the sentences based on the similarity of the pairs of words. Using *Kahina* text analytic tool, redundant sentences which are mostly word to word similar are merged or combined into one generic sentence and sentences which are related to well-known medical standards on diabetes diagnosis and treatments are discarded. Example of sentences filtered out are shown in Fig. 2. The remaining sentences are initially considered to support the judgment of new diabetes diagnosis or treatment and as consequence they are retained for further analysis.

file2928253 A value less than 6.5% does not exclude diabetes *diagnosed* using glucose tests.

file2928253 A value of less than 6.5% does not exclude diabetes *diagnosed* using glucose tests.

file1604865 Type 1 diabetes is *treated* with insulin , exercise , and a diabetic diet . Type 2 diabetes is first treated with weight reduction, a diabetic diet, and exercise.

file1604865 Type 1 diabetes is treated with insulin , exercise , and a diabetic diet . Type 2 diabetes is first *treated* with weight reduction, a diabetic diet, and exercise.

file1603650 Type 2 diabetes usually occurs slowly over time.

file1603650 Type 2 diabetes can also develop in people who are thin.

Fig 2. Sample of merged and filtered out sentences

3.2 Intuitive Prediction

The intuitive task focused on the prediction of the diabetes disease status and its association to new diagnosis and treatment based on implicit textual assertions. We used the same *Kahina* tool to process the remaining retrieved sentences (stemming, POS tagging and tokenisation of words) relying on a combination of term and few basic clinical inference rule-matching to extract new sentences related to diabetes diagnosis and treatment. We used larger combinations of diabetes related terms to search in the social network and research papers corpuses to find new association or concordance between the search terms. Examples of this includes “diabetes” and the words in the same sentence like “thirsty” and when digging down in the text source other words like “excessive hunger” and “fruity breath” as shown in Fig. 3 This could be taken as diabetes diagnoses are thirst, hunger and fruity breath, which are well known to characterize diabetes disease. The intuitive module consisted of two steps below.

Step #1: Candidate sentence identification. In this first intuitive step, the system identifies new potential evidence sentences in the social network and research papers corpuses, as shown in Fig. 3 below, by looking for any of the following three evidence types within the sentences (i) terms referring to the Diabetes disease symptoms (e.g., *thirst, frequent urine*), (ii) Important clinical facts or conditions related to the disease (e.g., *blurred vision, loss of weight*) and (iii) natural medications typically used to treat the disease and/or symptoms (e.g. *Chinese herbs and teas*).

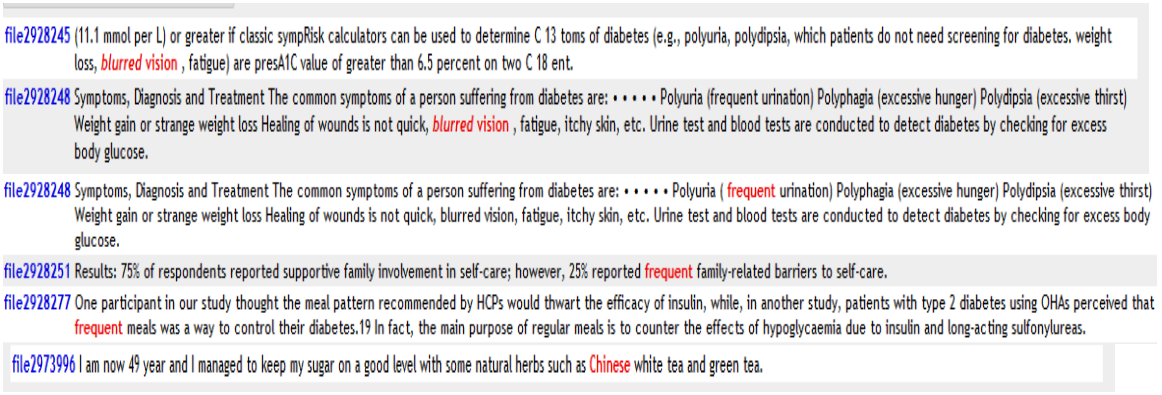

Fig. 3. Sample of Infered sentences

Step #2: Filtering Discovery Sentences. We use the same step 2 as in the descriptive prediction. So once the inferred sentences are selected as shown in Fig. 3 above, two filtering out processes are undertaken. *Kahina* tool implements basic sentence similarity algorithms by partitioning sentence into a list of tokens, then undertake POS tagging, and perform stemming words and finally identify similar sentences based on the similarity of the pairs of words. Using this *Kahina* tool redundant sentences which are mostly word to word similar are combined in one generic sentence. Also sentences which are related to well-known medical standards on diabetes diagnosis and treatments are discarded. The remaining sentences considered to support the judgment of new diabetes diagnosis or treatment are retained for further analysis.

3.3 Final Result integration.

The discovered set of sentences resulting from descriptive and intuitive predictions are integrated together. Similarly we used *Kahina* text analytic tool implements basic sentence similarity algorithms to compute the similarity of the sentences based on the similarity of the pairs of words and redundant sentences are combined in one generic sentence. Also sentences which are related to medical standards on diabetes diagnosis and treatments are discarded. The remaining sentences are initially considered to support the judgment of new diabetes diagnosis or treatment and as consequence retained for further analysis; a sample of such retained sentences are shown in Fig. 4 below.

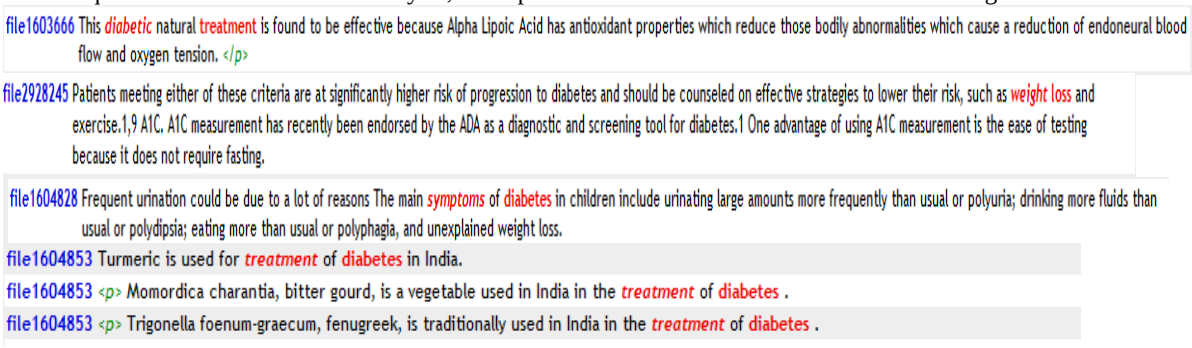

Fig. 4. Sample of final sentences retained for further analysis

4. Results and Conclusions

This paper has presented the use text mining techniques on the web and literature collected from the web and MEDLINE medical database. Many attempts were made on applying text mining functionalities to medical records, usually the ones obtained from MEDLINE database. The purpose here is to see if there could be any significance in

using social network of online stories of people where they share their accounts in their own words unlike the structured information obtained from medical databases and journals. For achieving the objective of this project, various methods and approaches were looked to obtain perspectives regarding this area of topic. Extensive research has been done in the area of text mining tools and text mining pre-processing approaches for the purpose of this project. Our *Kahina* text mining tool we used provides efficient text processing and analysis processes, which has given good results as shown below. The results of this research work will be put forward as hypothesis to medical experts and clinicians for further research and validations. It can also be added to an evolving ontology-based knowledge repository.

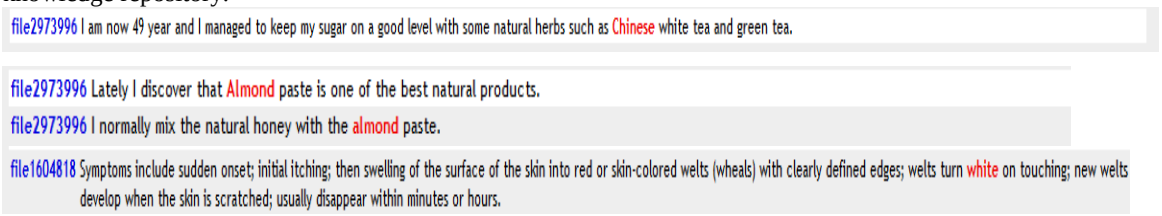

Fig. 5. Sample of final reslts after applying Kahina text processing

Acknowledgements

This research work is sponsored by the Research Incentive Fund (RIF No. 52014) in Zayed University, UAE.

References

1. J. Meikle, “Late diabetes diagnosis 'killing thousands early”. *The Guardian, [online] 11 June. Available at: < http://www.guardian.co.uk/society/2001/jun/11/health.voluntarysector>* [Accessed 25 March 2014].
2. E. Woolley, “Signs and Symptoms of Early Diabetes: A Detailed List to Help Catch Type 2 Diabetes Early”. *[online] (31 December 2011) Available at: <http://diabetes.about.com/od/ >*
3. S. Ananiadou, T. Ohta and K. Martin “Text Mining Supporting Search for Knowledge Discovery in Diabetes”. *Published online: 22 December 2012, Springer Science+Business Media* New York 2012
4. H. Yang, I. Spasic, J. A. KEANE, and G. Nenadic. “A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries” *Journal of the American Medical Informatics Association* Volume 16 Number 4 July / August 2009
5. M. Dittenbach, H. Berger D. Merll "Improving domain ontologies by mining semantics from text", APCCM '04 Proceedings of the first Asian-Pacific conference on Conceptual modelling, Darlinghurst, Australia, 2004. pp. 91-100.
6. M. Abulaish, L. Dey "Biological Ontology Enhancement with Fuzzy Relations: A Text-Mining Framework", Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, 2005. pp. 379-385.
7. S. Yin,Y. Qiu, J. Ge "Research and Realization of Text Mining Algorithm on Web", Computational Intelligence and Security Workshop, Harbin, December 15-19, 2007. pp. 413-416.
8. K.M. Sam and C.R. Chatwin"Ontology-based text-mining model for social network analysis", Management of Innovation and Technology (ICMIT), 2012 IEEE International Conference on, Sanur Bali, June 11-13, 2012. pp. 226-231.
9. Lee et al. “Bio-STEER: A Semantic Web workflow tool for Grid computing in the life sciences”, The International Journal of Grid Computing, Vol. 23, Issue 3, 2007.
10. A. Holzinger, K.M. Simonic, P. Yildirim, "Disease-disease relationships for rheumatic diseases: Web-Based Biomedical Textmining and knowledge discovery to assist medical decision making", Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual, Izmir, Turkey, 2012. pp. 573-580, 2012.
11. C. Vicient, D. Sanchez, A. Moreno “Ontology-Based Feature Extraction”, Web Intelligence and Intelligent Agent Technology (WI-LAT), 2011 IEEE/WIC/ACM International Conference on, Lyon, 2011. pp. 189-192.
12. A. Sharma, R. Swaminathan, Y. Hui "A Verb-Centric Approach for Relationship Extraction in Biomedical Text", Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, Pittsburgh, September 22-24, 2010. pp. 377-385.
13. L. Dey, and S.K.M Haque "Studying the effects of noisy text on text mining applications", AND '09 Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, New York, USA, 2009. pp. 107-114.
14. J. Clark “Text Mining and Scholarly” Publishing [pdf]. Available at: <http://www.publishingresearch.net/documents/PRCTextMiningandScholarlyPublishin Feb2013.pdf > [Accessed 03 April 2014].
15. M.A Hearst “Untangling Text Data Mining”. 37th Annual Meeting of the Association for Computational Linguistics, Proceedings of ACL'99. Available at: <http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html > [Accessed 03 April 2014].
16. D.R. Swanson “Fish Oil, Raynaud's syndrome, and undiscovered public knowledge”. Perspectives in Biology and Medicine, 30, 7-18, 1986.