International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 3568-3573
© Research India Publications. http://www.ripublication.com

# Disease-Treatment Relationship Extraction for Psoriasis from Online Healthcare Forums using NLP and Classification Techniques

**Ms. Mamatha Balipa**

*Department of Master of Computer Applications, NMAM Institute of Technology, Nitte, India.*

**Dr. Balasubramani R**

*Professor, Department of IS & E, NMAM Institute of Technology, Nitte, India.*

## Abstract

People consult a physician for a treatment for their ailments. But nowadays many people also search the web for health related issues. This paper describes the methodology using which treatments posted on online health care forums for the disease Psoriasis are extracted using NLP and classification techniques and given to the end user in a consolidated form.

**Keywords:** Psoriasis, Machine learning, Classification, Naive Bayes, Decision Tree, SVM

## INTRODUCTION

When people have ailments, they usually consult a physician for treatment for the disease. Nowadays the internet is considered as a knowledge base. Most of the people have access to the internet and the internet can be accessed using a desktop or any hand held device like the mobile phone. People also tend to share information over the internet, be it regarding a product or a medical condition. Even though the first step taken by people is to consult a physician in the case of any ailment, nowadays many people for various reasons consult medical resources available on the internet like health forums, discussion boards, blogs, etc., for information or treatments available for a disease, which are shared by other people. When a user is searching for treatments or solutions discussed online for a particular disease using a search engine, he she usually gets a list of links which the user has to traverse to get text that is posted by users where the treatment they have undergone for that particular disease and which has worked for them are described. To get this information, the user has to wade through a lot of unwanted information like more questions on the disease, treatments that have not worked, etc., The information is available in the form of text in natural language. Thus it is a challenge to extract relevant information from such text. This paper describes a method to extract treatments solutions posted by users for the disease Psoriasis online that have worked for them. Psoriasis is a type of auto immune skin disease for which there is no permanent solution found till date. Various types of treatments have worked for different people like Allopathic, Ayurveda, Homeopathy, etc., which have been discussed by the users online. Since the information is available in natural language, NLP techniques are applied to extract the relevant text.

## LITERATURE

### *A. Health related information on the web*

The popularity of the social media is growing and they provide opportunities to study interactions among humans and their experiences. In the last ten years there is an increase in the consumption of health information available online. Online health information act as a rich resource for health researchers too. Pew Internet reports in 2011 in The Social Life of Health Information, that 80% of the people who have used the internet have referred a website for health related information [13] and 59% of people among them have searched for information related to certain medical issues. An online survey conducted by Angus Reid on 1,010 Canadian people who were randomly selected, showed that 89%of the people consulted the web to research health related issues and symptoms [3]. Similarly, it was found by Pew Research Internet Project that in 2013, 59% of Americans searched for health related issues online [3]. Health related information extracted from social media sites like Facebook and Twitter have been used in research by certain organizations[3]. Hariprasad Sampathkumar et al. have extracted drug side-effects from messages posted on the web which help in post-marketing drug surveillance[10]. Leaman et al. [6] in their research work created a dictionary of terms related to adverse effects of certain drugs and used a sliding window technique to identify adverse drug reactions in messages posted by users in online health forums. Li [7] used statistical techniques to analyse messages in sites where reviews on drugs are posted, to identify relations between certain class of drugs and some of the health disorders, which can be backed by existing research literature. Similarly Wu et al.[15] created UDWarning, that identifies side effects of drugs that were not recognized earlier. They have used co-occurrence statistics to identify the relevance of a web page that has messages posted about the side effects of various medications. Liu et al. [8] proposed a framework called AZDrugMiner that used statistical techniques to retrieve side effects of various medications from messages posted online by patients. Chee et al.[2] used natural language processing to perform sentiment analysis and they used a combination of classifiers to find posts on drugs that are monitored by the FDA. They used messages posted by users on Health and Wellness Yahoo Groups. Bian et al.[1] used Support Vector Machine(SVM) for classification and UMLS meta thesaurus to analyze textual and semantic fea- tures for the purpose of mining adverse drug reactions. Yang et al.[16] used Proportional Reporting Ratios and association rule mining to find relations between the ADRs(Adverse Drug Reactions) and the drugs used from user messages posted on health forums. V.G.Vinod Vydiswaran, PhD, et all in their research, proposed a pattern-based approach to mine the text from Wikipedia for dentifying realtions between commonly used terms for a health issue by laymen and terms for the same used by professionals.[14].

### *B. Online healthcare forums*

Patients share information on social media and healthcare forums like psoriasis-association.org.uk, www.healingwell.com, MedHelp.org [5] and Health-Boards.com [5] about their health issues, medications they have used or treatments they have undergone. Messages posted on healthcare forums on the internet generally have spelling and grammer errors and words which are unclear and vague. This will not be the case with information from medical and health centres. It is a challenge extracting useful information from such forums. But the access to unrestricted and first-hand information provided by patients has many researchers exploring the possibility of mining useful health information from such online forums. MedHelp is one such well-known online forums where patients discuss health issues. The dataset has nearly thirty million posts by nearly a million unique users, and consists of nearly 450 million words. Many researchrs have used the dataset in their research endeavors[5].

*1) Natural Language Processing and Text Mining:* NLP makes use of computational techniques and linguistic concepts to analyse natural language and speech. Parsing of the text and Parts-of-Speech tagging are one of the common methods used in NLP. Nouns, verbs, adjectives, adverbs, etc., are identified using lexical databases like VerbNet and WordNet. Popescu and Etzioni used NLP, MINIPAR parser, relationship between words to infer nouns using adjectives, and also WordNet for information about synonyms and antonyms. The resultant structures got by parsing the text were used to mine expected realtions using hand coded rules. Adaptation of Support Vector Machines (SVMs) were used by Bunescu and Mooney for extracting relations, and also to compare NLP and non-NLP techniques. Documents are usually treated as unstructured bag of words and SVMs are widely used for text classification. In bag of words method, only which words occur and their counts matter, not their structural information or positions. Mustafaraj et al. also used statistical methods along with parsing to classify text. They classify text by combining three different classifiers. They also use hand coded rules for extracting lexical relations between words and off-the-shelf POS taggers. They have also utilized VerbNet and FrameNet to discover verbs that are important and relationships between words to incorporate into their knowledge base and used them for analysis off-line. Finally, they have used dynamic learning, a comparatively new statistical method applied in text mining to increase the effeciency of their classifiers. Marchisio et al. solely use NLP methods, developing a parser and a technique to index and simplify complicated parse trees that capture elementary linguistic associations among words. Hariprasad Sampathkumar et al[10] extracted side- effects of drugs from messages posted on health message boards by making use of Hidden Markov Model(HMM) for classification. Messages posted on one of the health forms were annotated and maade use of as a data set to train and validate the HMM based Text Mining system. A ten fold cross validation performed on the annotated dataset using HMM classfier provided an F-Score of 0.76. Jang et al.[5] mined symptoms, treatments and performance information in clinical documents, using a semantic tagger which also made use of the HMM classifier. The documents contained a combination of Korean and English words. Wang et al. conducted a study on the viability of using statistical and NLP techniques on information available in Electronic Health Records to support the use of computers in Pharmacovigilance. Warrer et al. have reviewed text mining techniques on electronic patient records to identify ADRs. Sohn et al.[11] applied rule-based methods for extracting side-effects of various medications from clinical sources related to psychiatry patients.

### *C. Entity Relationship Classification Techniques*

The task of Relationship Classification is used to identify the types of relationships between entities. The different approaches for relationship extraction between entities are:

*1) Co-Occurrences Analysis:* The co-occurrences approach is based on knowledge of the relationships between words, their structure and context. The approach provides good level of recall but low precision. Jenssen et al. and Stapley and Benoit [12] applied the technique on abstracts from Medline.

*2) Rule based approaches:* Rule based approach has been commonly used in relationship extraction. In this approach either syntactic (that is Parts of Speech) or semantic information like fixed patterns containing certain words indicating relations are used. Only drawback is that more of human effort is required. Rule based approach can be semi-automatic or manual. Rule based approach give good precision but sometimes low recall.

*3) Statistical Methods:* Statistical approaches are also used along with NLP on annotated corpora. Statistical approaches do not require huge training data. Learning algorithms automatically extract rules. In this approach the text is considered as a bag-of-words. Some researchers use the approach along with other information like POS.

## EXPERIMENTS AND RESULTS

Information retrieval was performed using crawlers to extract messages from sources like psoriasis-association.org.uk, healingwell.com, MedHelp.org [5] and HealthBoards.com [5]. The search engineto do the same was developed JSoup API[4], a Java HTML parser library and Apache Lucene[9]. About 2000 posts were collected from psoriasis-association.org.uk, healingwell.com, MedHelp.org [5] and HealthBoards.com [5]. The Text processing part of the system was used to extract the text messages from the document collection. The text extracted was preprocessed and then transformed into a form which can then be used by the information extraction part of the system. The information extraction part of the system uses a pipeline of NLP techniques to process the text.

The Information Extraction part of the system is used for POS tagging as well as to identify the entities of interest like the treatment names. The Relation Extraction module identifies the relationship between the Named Entities and a combination of techniques including rule-based, natural language processing techniques, use of classfiers, word context, co-occurance analysis and statistical have been used for this purpose. The author has applied supervised machine learning approaches like Naive Bayes, Decision Tree, Support Vector Machine and Logistic Regression to classify the comment as a treatment for the disease or not a treatment.

### *A. Healingwell.com, MedHelp.org and HealthBoards.com*

Healingwell.com, MedHelp.org [5] and Health-Boards.com.com [5] are health forums having multiple threads discussing issues regarding Psoriasis. Users discuss treatments they have undergone that have not worked, treatments that have worked, post questions, food that aggravate the symptoms or are the cause for the disease, food that give relief from the symptoms and all the issues pertaining to the disease. A message may consist of single sentence or multiple sentences. Since the messages are free flow of text in English, the text needs to be transformed into a form which can be processed as well as since the messages are in natural language, NLP techniques need to be applied to extract information from the text. Since the data used is extracted from online healthcare forums, the system utilizes the features of Big Data. Healthcare message boards available online provide huge volume of latest and raw data which can be used to mine useful information.

### *B. Detailed Description*

Information retrieval: The first step is information retrieval. A search engine was developed that will search and download all the pages from the web pertaining to the disease Psoriasis. To check the relevance of the page, a threshold value for the count of the number of times the word Psoriasis occurs in the page is maintained. The search engine was developed using Apache Lucene and Jsoup API. Using JSoup API, individual comments from the online users in the page are extracted and a corpus of text containing the comments is created.

Topic detection: It is ensured that the topic of discussion in the page is about Psoriasis by using Latent Dirichlet Allocation (LDA) model. For this, first the text is normalized by eliminating stop words, punctuation symbols and lemmatizing the text. A term dictionary of the text and Document Term Matrix is created. Finally the topic of discussion in the text is arrived at by applying the LDA model on the document term matrix. Feature extraction: The comments in the corpus are categorized into solutions and non solutions. An algorithm is developed and implemented to extract features that identifies a particular text as solution or treatment for the disease Psoriasis. The algorithm works as follows. First the corpus of comments is read. The corpus has comments that suggest solutions as well as comments that are not solutions. The comments present in the corpus are manually categorized as solutions and non- solutions. All the text present in the solution category are extracted. Using regular expression all the alphabets, digits and exclamatory marks are extracted from the text and other unnecessary symbols removed. The text extracted using reg- ular expression is converted to lowercase. The text is further tokenized and the list of words present in the text is retrieved. The most commonly occurring words in comments categorized as solution is found using frequency distribution. The list of unique unigrams are listed from comments categorized as solutions. The unigrams that are not available in the English dictionary and chat dictionary are found and added to rare words list. The unigrams are tagged using POS tagging and all unigrams following the pattern unigram/NUM mg of or precedes the pattern unigram/NUM mg are added to rare words list. They are stored in a file. The most commonly occurring rare words are found using frequency distribution. The most commonly occurring bigrams and trigrams occurring in the comments categorized as solutions are found. Each comment in the corpus is tagged as solution or non-solution. The list of tagged text are shuffled. Feature sets identifying a comment as either solution or non-solution is extracted. The algorithm to extract the feature sets is as follows. In each comment, the presence of the most common trigrams identified in the solutions text is checked and a score is given for the existence of each trigram in the comment. The list of words in the comment that are not stop words are extracted and a score given. Unigrams present in the comment are extracted and their presence in the list of common unigrams found in solutions is checked. In addition to the above, the exis- tence of word patterns like grateful, relief, improvement, best, treatment, disappeared, believe, appeared, completely, healed, worked, discovery, purely, chance, happy, better, quickly, smooth, clear, cleared, gone, etc., are also checked and weights assigned to them according to their gravity. The existence of bigram patterns like, cannot believe, no longer, application healed, also worked, little slower, treatment brought, great relief, have any, really soothing, huge plus, best solution, best results, completely clear, truly amazing, recommend it, really good, etc., are checked in the comment and weights assigned to them according to their gravity. Text in the comment that has a numeric value followed by the string mg is extracted. It is found that the comments that have a pattern like 200 mg, etc., will usually be a comment containing a solution. To identify a number, Parts of Speech tagging is done for the tokens in the comment and all tokens with the POS tag NUM is identified and also the count of the number of times the NUM POS tag occurs in the comment is counted. It is also found that the comments that contain more nouns are usually solutions. So the count of the number terms with POS tag NOUN is also found in the comment. Also it is found that the text that contain exclamatory marks (!) are also usually solutions. A count of the number of times ! occurs is also found. The words not occurring in the English dictionary like panchakarma, ayurveda, and other medicine names, that are occurring in the comment are extracted. The context of each word is found using concordance. Once again the occurrence of the common unigrams, bigrams and trigrams and the patterns like numeric values, the string mg and exclamatory marks are checked in the context. For each comment, the above feature sets are extracted and the features categorized as belonging to solutions or non- solutions. The feature sets are divided into 70 % training set and 30 % testing set. A ten fold cross validation is performed. The Naive Bayes classifier is trained with 70% of the data and the remaining 30% is used for testing. The Naive Bayes classifier provides an accuracy of 98 %. Similarly Decision tree is applied on the same data sets and an accuracy of 94%

**Table III:** Precision, Recall, F Measure for Naive Bayes

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 1.0 | 0.8 | 0.888888888889 |
| Pos | 0.923076923077 | 1.0 | 0.96 |

*1). Decision Tree classifier:* Decision Tree classifier pro- vides an accuracy of 88.2352941176

**Table IV:** Confusion Matrix For Decision Tree

| | Neg | Pos |
|---|---|---|
| Neg | 1.0 | 2 |
| Pos | 0 | 14 |

**Table V:** Precision, Recall, F Measure for Decision Tree

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 1.0 | 0.333333333333 | 0.5 |
| Pos | 0.875 | 1.0 | 0.933333333333 |

*2) SVM classifier:* SVM classifier provides an accuracy of is got. SVM is also applied on the same data sets and an88.2352941176 accuracy of 88% is got.

*C. Performance Analysis of Classification Techniques*

*1) Naive Bayes classifier:* Naive Bayes classifier provides an accuracy of 94.1176470588

**Table VI:** Confusion Matrix for SVM

| | Neg | Pos |
|---|---|---|
| Neg | 10 | 1 |
| Pos | 1 | 5 |

**Table I:** Naive Bayes Most Informative Features

| | |
|---|---|
| contains(clear) = True | pos : neg = 7.4 : 1.0 |
| contains(quickly) = False | neg : pos = 4.8 : 1.0 |
| contains(believe) = False | neg : pos = 4.8 : 1.0 |
| contains(appeared) = False | neg : pos = 4.8 : 1.0 |
| contains(soothing) = False | neg : pos = 4.8 : 1.0 |
| contains(great) = False | neg : pos = 4.8 : 1.0 |
| contains(purely) = False | neg : pos = 4.8 : 1.0 |
| contains(clear) = False | neg : pos = 4.8 : 1.0 |
| contains(cannot believe)= False | neg : pos = 4.8 : 1.0 |
| contains(grateful) = False | neg : pos = 4.8 : 1.0 |

**Table VII:** Precision, Recall, F Measure For Svm

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 0.909090909091 | 0.909090909091 | 0.909090909091 |
| Pos | 0.833333333333 | 0.833333333333 | 0.833333333333 |

*4) SVC classifier:* SVC classifier provides an accuracy of 52.94117647058824

**Table VIII:** Confusion Matrix for SVC

| | Neg | Pos |
|---|---|---|
| Neg | 3 | 4 |
| Pos | 4 | 6 |

**Table II:** Confusion Matrix for Naive Bayes

| | Neg | Pos |
|---|---|---|
| Neg | 4 | 1 |
| Pos | 0 | 12 |

**Table IX:** Precision, Recall, F Measure For Svc

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 0.429 | 0.429 | 0.429 |
| Pos | 0.6 | 0.6 | 0.6 |

*5) LinearSVC classifier:* LinearSVC classifier provides an accuracy of 82.35294117647058

**Table X:** Confusion Matrix For Linearsvc

| | Neg | Pos |
|---|---|---|
| Neg | 9 | 0 |
| Pos | 3 | 5 |

**Table XI:** Precision, Recall, F Measure for Linearsvc

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 0.75 | 1.0 | 0.8571428571428571 |
| Pos | 1.0 | 0.625 | 0.7692307692307693 |

*6) NuSVC classifier:* NuSVC classifier provides an accuracy of 70.58823529411765

**Table XII:** Confusion Matrix for NUSVC

| | Neg | Pos |
|---|---|---|
| Neg | 6 | 2 |
| Pos | 3 | 6 |

**Table XIII:** Precision, Recall, F Measure for NUSVC

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 0.67 | 0.75 | 0.71 |
| Pos | 0.75 | 0.67 | 0.71 |

*7) Logistic Regression classifier:* Logistic Regression classifier provides an accuracy of 100.0

**Table XIV:** Confusion Matrix for Logistic Regression

| | Neg | Pos |
|---|---|---|
| Neg | 6 | 0 |
| Pos | 0 | 11 |

**Table XV:** Precision, Recall, F Measure For Logistic Regression

| Label | Precision | Recall | F Measure |
|---|---|---|---|
| Neg | 1.0 | 1.0 | 1.0 |
| Pos | 1.0 | 1.0 | 1.0 |

## LIMITATIONS

The limitations of this approach is that, since the messages are posted by average users, there may be noise, inaccurate and exaggerated information with spelling mistakes. Mining such information may lead to false positives. But the volume of the data may help in solving this problem. Repeatedly occurring treatments can be considered as true positives. Some of the messages that are posted may be to promote certain drugs or products. So further investigation by medical experts may be required.

## FUTURE WORK

The Naive Bayes classifier and Logistic Regression are found to be promising in extracting disease-treatment rela- tionships for the disease Psoriasis from messages posted on various internet message boards. The data mined from these sources is utilized to provide statistics regarding the success rates of different types of treatments and the medications used in those treatments. The extracted data can be represented using ontologies which can then be used for querying unique and more accurate information. The authors would also like to explore other Machine Learning algorithms like the Neural Network, HMM classifier and other deep learning algorithms.

## CONCLUSION

In this paper an effort has been made to describe the method using which user comments where solutions or treatments for the disease Psoriasis that have worked for users and posted online in health forums are extracted and displayed in consolidated form. This work is in no way recommending any treatment or medicine. It only extracts the treatments that provide solutions or look promising by analyzing online user comments and gives it to the end user. The data can be used for a comparison study of the different types of treatments available for the disease Psoriasis.

## REFERENCES

[1] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.

[2] Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 217. American Medical Informatics Association, 2011.

[3] Diana L Gustafson and Claire F Woodworth. Methodological and ethical issues in research using social media: a metamethod of human papillomavirus vaccine studies. *BMC medical research methodology*, 14(1):127, 2014.

[4] Jonathan Hedley et al. jsoup: Java html parser, 2015. *Website (https://jsoup. org/).*

[5] Hyeju Jang, Sa-Kwang Song, and Sung-Hyon Myaeng. Text mining for medical documents using a hidden markov model. In *AIRS*, pages 553–559. Springer, 2006.

[6] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125. Association for Computational Linguistics, 2010.

[7] Yueyang Alice Li. *Medical data mining: Improving information acces- sibility using online patient drug*

*reviews*. PhD thesis, Massachusetts Institute of Technology, 2011.

[8] Xiao Liu and Hsinchun Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *International Conference on Smart Health*, pages 134–150. Springer, 2013.

[9] Apache Lucene. Apache lucene core, 2013.

[10] Hariprasad Sampathkumar, Xue-wen Chen, and Bo Luo. Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC medical informatics and decision making*, 14(1):91, 2014.

[11] Sunghwan Sohn, Jean-Pierre A Kocher, Christopher G Chute, and Guergana K Savova. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, 18(Supplement 1):i144–i149, 2011.

[12] Benjamin J Stapley and Gerry Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Biocomputing 2000*, pages 529–540. World Scientific, 1999.

[13] V. G Vinod Vydiswaran, Qiaozhu Mei, David A. Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014:1150–1159, 2014.

[14] VG Vinod Vydiswaran, Yang Liu, Kai Zheng, David A Hanauer, and Qiaozhu Mei. User-created groups in health forums: What makes them special? In *ICWSM*, 2014.

[15] Hao Wu, Hui Fang, and Steven J Stanhope. An early warning system for unrecognized drug side effects discovery. In *Proceedings of the 21st International Conference on World Wide Web*, pages 437–440. ACM, 2012.

[16] Christopher C Yang, Haodong Yang, Ling Jiang, and Mi Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 33– 40. ACM, 2012.