

HYBRID APPROACH FOR DETECTING THE CAUSE FOR THE SPREAD OF WATERBORNE DISEASES FROM THE SOCIAL MEDIA SERVICES USING MAPREDUCE

R. Sriranjani,

K. Shenbagalakshmi,

S. Pradeepa

School of Computing, SASTRA Deemed University, India

Abstract-waterborne disease is the major threat to India in the 21st century. Thus, requiring a system which detects the major cause for the spread of waterborne diseases, and take a preventive measure before any major loss. The following methodology provides an approach for detecting the spread of waterborne diseases based on data collected from social media services. The system not only detects the cause for the disease, it also detects the probability of various waterborne diseases, given the symptom based on trend. The model overcomes the drawback of processing data in the single node environment. It uses the Hadoop map reduce framework to process the data which makes it simple to process huge volume of data. The main advantage of using social media is that it provides data that is very close to real time.

I. Introduction

The waterborne disease is transmitted by drinking the water, which is contaminated by pathogenic microorganisms. In most of the developing countries like India, there are increasingly large amount childhood deaths due to improper water hygiene and sanitation.

Hence, it is essential to be aware of the symptoms for the spread of the disease [1]. In the recent years social media is becoming a major part in everyone's living. This is attracting a large amount of research topic based on social media data sources [2]. In the proposed system, data that is posted related to waterborne diseases are collected from various social media services like Twitter, Wikipedia and so on.

The proposed approach can be used for detecting the probability of various causes for the spread of waterborne diseases. The system initially finds the probability of various symptoms causing the disease using Naive Bayesian algorithm [3] and then extends it to detect the probability of waterborne disease.

Most existing research limited the work till finding the cause of the disease in the single node environment [12]. This leads to a difficult situation when the size of the data increases gradually. Moreover, the existing approaches do not calculate the probability of the disease, given the symptom.

Since, the system implements Naive Bayesian algorithm under the map reduce framework[4] to calculate the probabilities, it becomes easy to work with large amount of overwhelming data collected from various social media sources. The approach also provides a method to detect the disease from the given set of symptoms based on the calculated probabilities. This can help us to take preventive measures before the disease outbreak.

The salient features of the proposed system are:

- Ability to handle large amount of overwhelming social media data.
- Accurate probability calculation using Naive Bayesian algorithm.
- Efficient and scalable for practical use.
- Ability to capture real-time events from social media services.

II. Related Work

In the recent years, there is various research papers that are discussing about several diseases connected with social media services.

The work done by Vinay Kumar Jain and Shishir Kumar provides an approach to detect the prevalence of mosquito-borne diseases using social media [5]. The system uses information collected from Geographical Information System to predict the location of the disease in India.

Similarly, the work proposed by Georgiana Ifrim, Bichen Shi and Igor Brigadir uses filtering and clustering method for event detection in Twitter [6]. This technique initially performs hierarchical clustering of tweets along with aggressive filtering based on length and structure. Finally, the results are ranked based on the outcome of clustering.

To improve the accuracy of classification Tajinder Singh and Madhu Kumari provided an approach for pre-processing of text in the sentiment analysis performed in Twitter [7]. This checks the significance of the slang word based on bindings of the slang words on other coexisting words.

Another approach that is provided by Aron Culotta [8] provides regression based modelling to discover correlation between influenza related tweets. This approach finds the correlation between various tweets to find relationship among different tweets.

Another work provided by Vinay Kumar Jain, Shishir Kumar [9] uses the data collected from social media to detect the disease outbreak. Further, the data is categorized into various parameters to collect relevant information related to swine flu.

Similarly, the methodology proposed by Vasileios Lampos, Nello Cristianini [10] provides a technique for searching messages related to symptoms and convert the information into a flu-score.

Another work proposed by Jon Parker, Yifang Wei and et al [11] provides an approach for tracking health conditions using Twitter. Using the fluctuations in medical-related posts over various time series is used to detect shifts in public health conditions.

Based on the related work, we provide an approach for detecting the cause of the spread of the waterborne disease.

The topics that are discussed in this paper are arranged as follows:

In section 1, we discuss about various waterborne diseases. Then, in section 2, we provide an insight into system architecture. The various pre-processing steps performed over the collected data set are explained in section 3. In section 4, we discuss about the algorithm that is used to find the probability of various symptoms of waterborne diseases. Further, in section 5, the process of performance evaluation is discussed. Finally, section 6 provides the conclusion of our approach.

The impact of waterborne disease (Malaria) on people can be inferred from the data collected by the Ministry of Health.

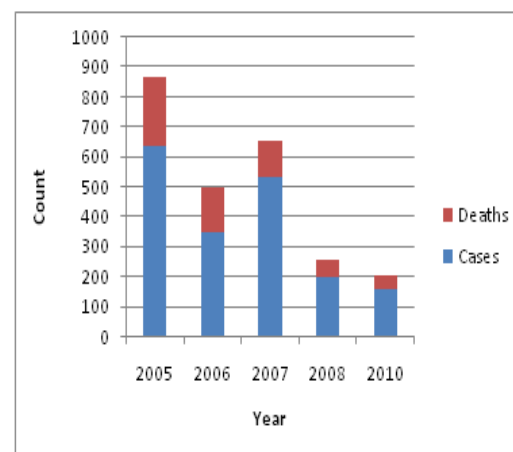


Fig 1. Bar chart depicts Malaria incidence in Chennai from 2005 to 2010.

1. Waterborne Diseases

a. Malaria:

Malaria is most commonly transmitted by an infected female Anopheles mosquito by introducing the parasites into person's blood through mosquito bite. The infection can be viewed through symptoms after 8-25 days of mosquito bite. The common symptoms are chill, fever, headache, and vomiting. Mosquito bites can be prevented by using insect repellent, covering your arms and legs.

b. Typhoid:

The infection is caused by Salmonella Typhi. It usually spreads through ingestion of contaminated food or water. The symptoms may last weeks or months. It includes high fever up to 104 degrees Fahrenheit over several days. Additionally abdominal pain, weakness, headaches and constipation also occur.

c. Cholera:

Cholera is an acute diarrheal infection. Ingestion of food or water contaminated with the bacterium Vibrio cholera is major cause for the infection. Symptoms occur after 12 hours and 5 days for a person after ingesting contaminated food or water. A large amount of watery diarrhea that lasts a few days is the main symptom. Preventive measures of cholera include drinking clean water and practicing good oral hygiene.

d. Diarrhea:

Diarrhea infections are often acquired from food or water that has been contaminated by feces, or directly from another person who is infected. The symptoms lasts for a few days and result in dehydration due to fluid loss. Infection of the intestines due to a virus, bacteria, or parasite is the common cause. Clean drinking water and improved sanitation prevents disease.

2. System Architecture

The objective of our system is to find the probability of the various symptoms causing the waterborne disease. It also provides probability of a disease, given a symptom. Our model is based on the data collected from various social media services like Twitter, Wikipedia and so on.

Initially, the data that is posted related to waterborne diseases are collected over a time span of four weeks from Feb-01 to Feb-29. Then, various Pre-processing tasks are performed over the collected data to remove unnecessary and noisy data.

Then, the processed data contains only relevant and useful information. This information is used as an input for the mapper task. The information passes through two different mapper where each one performs a different job. The output of the mapper is generated as a (key, value) pair. Some of the symptoms of waterborne diseases are listed below:

Symptoms: Fever, Headache, Diarrhea, Vomiting, Stomach ache, Weight loss, Dehydration, Lack of appetite, Cough, Chills, Extreme thirst, Muscle and joint pain, Low levels of blood platelets.

The output from the mapper is passed to the reducer task. Here, the actual implementation of Naive Bayesian algorithm is done. The probability of various symptoms causing the waterborne disease is calculated and it is attached as a value to the key.

The work is also extended to predict the disease, given any symptom, using previously calculated probabilities. This can be performed using Bayes' Theorem. The theorem can be stated as follows:

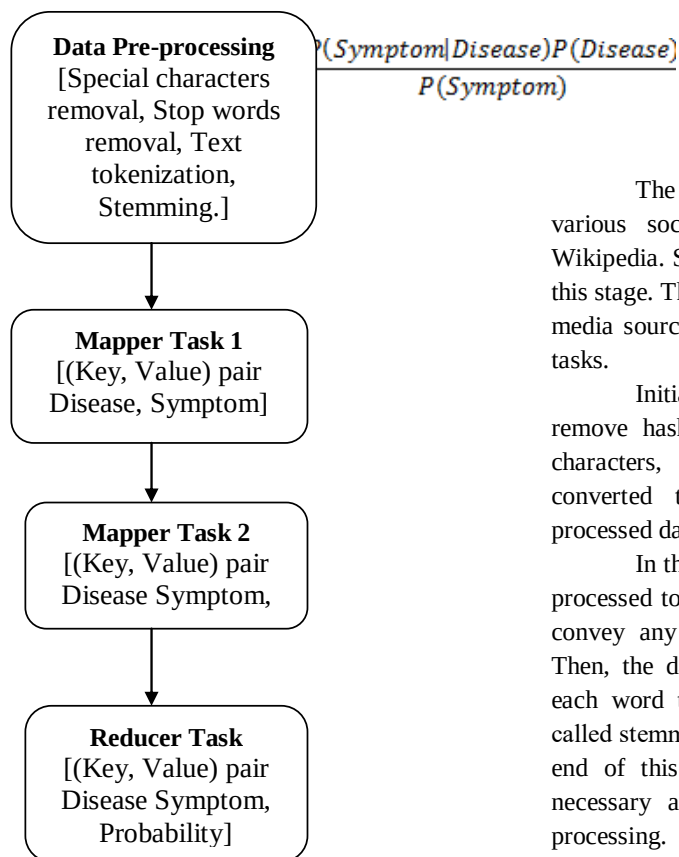


Fig 3. Depicts the detailed view of the system architecture.

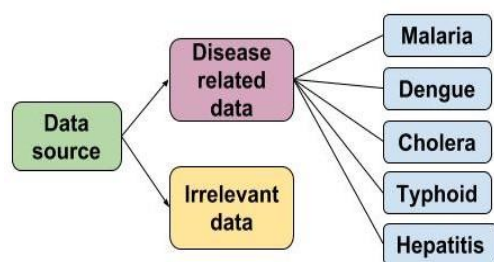


Fig 4. Depicts the classification of data source

3. Data Preprocessing

The data for our work is collected from various social media services like Twitter, Wikipedia. Several unwanted data is removed at this stage. The post collected from various social media sources undergoes various preprocessing tasks.

Initially, the data are processed to remove hash tags, punctuations, other special characters, urls, numbers and each word is converted to its lower case. Further, the processed data is tokenized by white spaces.

In the next step, the tokenized words are processed to remove stop words, as they do not convey any useful information for our work. Then, the data is further processed to convert each word to its base form by the technique called stemming using Porter's algorithm. At the end of this step, the data set contains only necessary and useful information for further processing.

The various processing performed at this step can be summarized as follows:

- Remove special characters.
- Convert to lower case.
- Text tokenization.
- Perform stemming.
- Stop words removal.

4. Probability Calculation

The probability of several symptoms related to waterborne diseases is calculated after various stages of processing. Since, the model works under the Hadoop framework, it has mapper and reducer phase.

a. Mapper Phase:

There are two mapper task involved at this phase. In the first phase of processing, the input data is converted into (key, value) pairs, where key denotes the waterborne disease and the value denotes its associated symptom. This, generated output is passed as an input to the second mapper task, which calculates the count of each (key, value) pair and assigns it as a value.

The input data received after preprocessing is distributed across different data nodes. The name node which coordinates the mapper task to associate the symptom of each disease in the phase 1. In the second phase, the name node is further used to calculate the count of each symptom and attaches it as the value to the (key, value) pair.

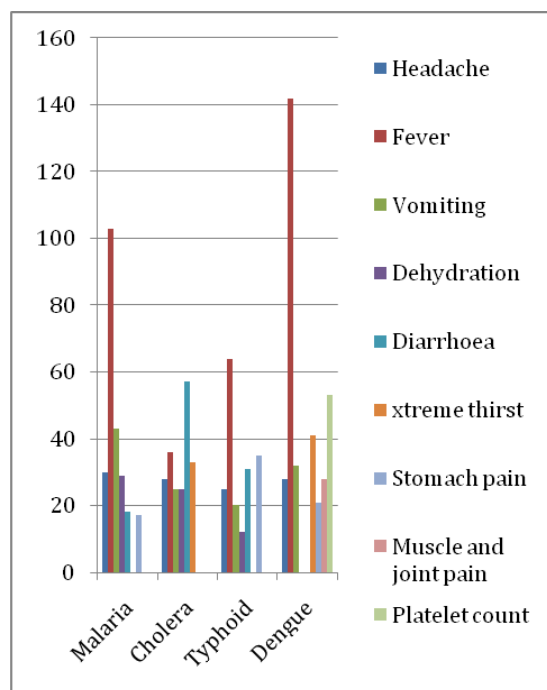


Fig 5. Pictorial representation of reducer output

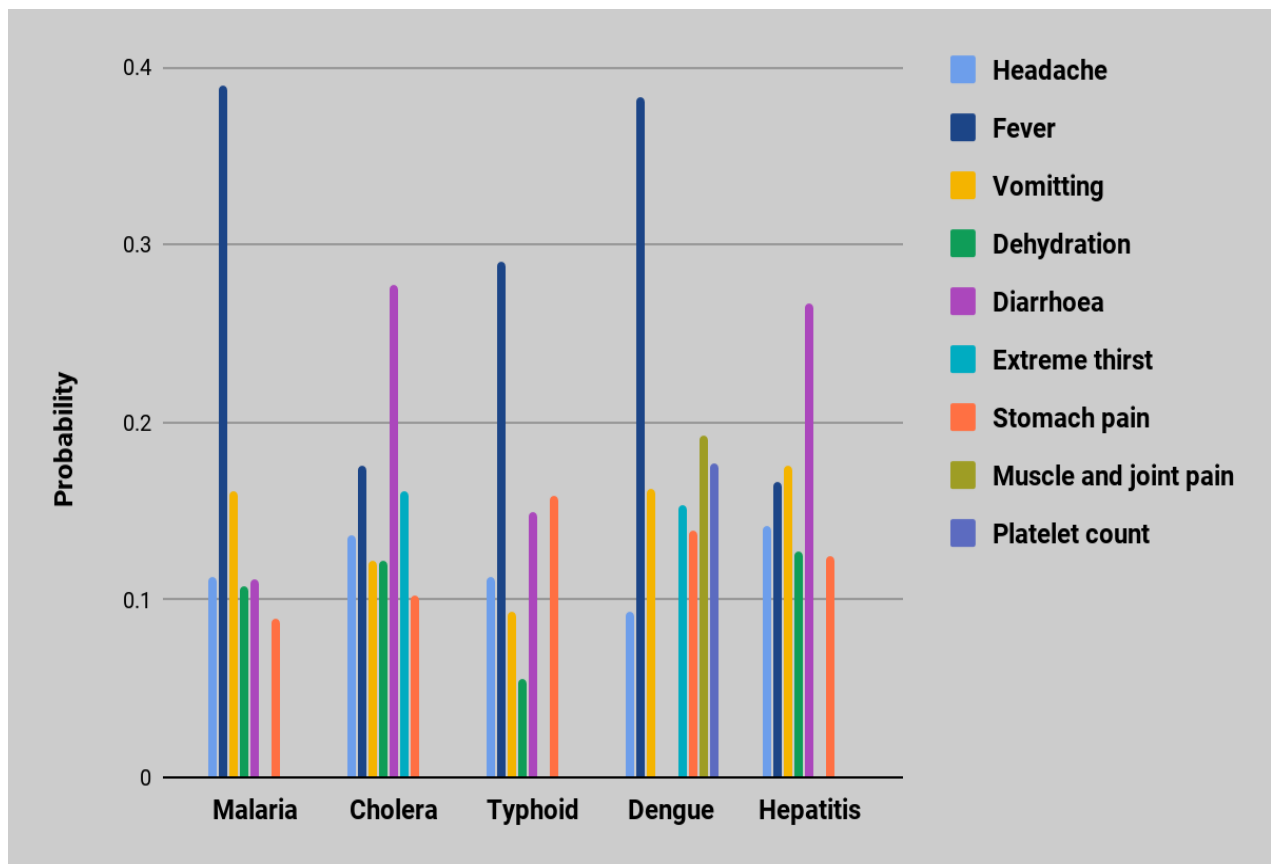


Fig 6. Pictorial representation of probability output.

b. Reducer Phase:

The actual implementation of Naive Bayesian algorithm is carried in this phase. In this phase, the output of the mapper phase is taken as input and aggregates the count of various symptoms. Then, the probability of each waterborne disease together with its symptom is calculated using Bayes' theorem [3] as stated in section 2. This probability is used to predict the major symptom for the spread of a particular waterborne disease.

c. Algorithm:

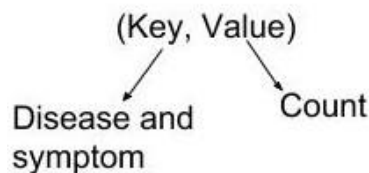
Step 1: Collect data related to waterborne diseases from social media services.

Step 2: Remove special characters and convert long sentences into tokens.

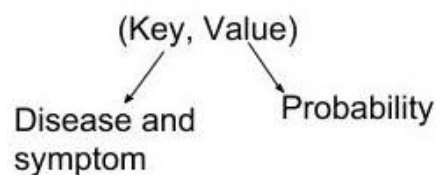
Step 3: Remove stemming words and convert every word to its base form.

Step 4: Remove stop words from previous input.

Step 5: Pass the preprocessed data to mapper. This produces a (key, value) pair. Here, key represents disease and its associated symptom and value represents its count.



Step 6: Pass the mapper output to the reducer. This produces a (key, value) pair. Here, key represents disease and its associated symptom and value represents its probability.



Here, the probability is calculated using Naive Bayesian algorithm.

5. Performance Evaluation

The output obtained from the reducer task when executed in map reduce environment can be visualized as represented in fig. 4 and 5. The obtained result coincides with the data collected by Ministry of health. The preprocessed data is stored in Hadoop Distributed File System (HDFS) which provides the advantage of scalability. There are two name nodes that are used for our work. It can be increased to improve the efficiency. The total time spent by map task is 13185 ms and the total time spent by reduce task is 3667 ms. This

approach provides an insight into the various symptoms of waterborne diseases. Thus, helping the common people and government officials to take rapid action to prevent major loss. The various symptoms of Typhoid obtained from the output are listed below:

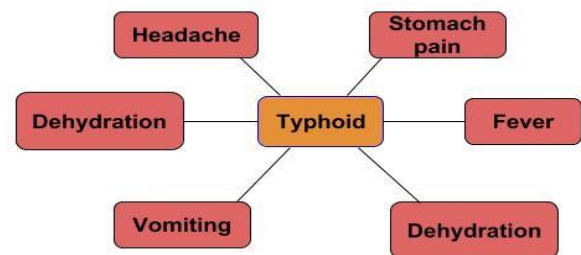


Fig 7. Depicts the various symptoms of Typhoid.

6. Conclusion

The various techniques discussed in earlier research papers offer various limitations. Traditional approaches induce scalability issues and time delays. The proposed methodology overcomes the disadvantages and improves the efficiency. An efficient symptom probability prediction system has been conveyed in this paper. A system that calculates the probability of various waterborne related symptom is proposed, which clearly improves the understanding of the common people towards the disease and also helps doctors for initial diagnosis.

The provided approach can be extended further to any domain. The approach can be modified to include location related information to precisely calculate the probability for further accuracy.

References

- [1] Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1), 39-45.
- [2] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 492-499). IEEE Computer Society.
- [3]https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- [4] <https://en.wikipedia.org/wiki/MapReduce>
- [5] Jain, V. K., & Kumar, S. (2017). Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science*.
- [6] Ifrim, G., Shi, B., & Brigadir, I. (2014, April). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW)*, Seoul, Korea, 8 April 2014. ACM.
- [7] Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, 89, 549-554.
- [8] Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115-122). ACM.
- [9] Jain, V. K., & Kumar, S. (2015). An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter. *Procedia Computer Science*, 70, 801-807.
- [10] Lampos, V., & Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (pp. 411-416). IEEE.
- [11] Parker, J., Wei, Y., Yates, A., Frieder, O., & Goharian, N. (2013, August). A framework for detecting public health trends with twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 556-563). ACM.
- [12] Chew, C. (2010). *Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak* (Doctoral dissertation).

