Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Topical Mining of malaria Using Social Media. A Text Mining Approach

James Boit
Dakota State University
James.boit@trojans.dsu.edu

Omar El-Gayar
Dakota State University
Omar.El-Gayar@dsu.edu

Abstract

Malaria is a life-threatening parasitic disease, common in subtropical and tropical climates caused by mosquitoes. Each year, several hundred thousand of people die from malaria infections. However, with the rapid growth, popularity and global reach of social media usage, a myriad of opportunities arises for extracting opinions and discourses on various topics and issues.

This research examines the public discourse, trends and emergent themes surrounding malaria discussion. We query Twitter corpus leveraging text mining algorithms to extract and analyze topical themes. Further, to investigate these dynamics, we use Crimson social media analytics software to analyze topical emergent themes and monitor malaria trends.

The findings reveal the discovery of pertinent topics and themes regarding malaria discourses. The implications include shedding insights to public health officials on sentiments and opinions shaping public discourse on malaria epidemic. The multi-dimensional analysis of data provides directions for future research and informs public policy decisions.

1. Introduction

According to the World Health Organization (WHO) [1], over 400,000 people die of malaria and a further estimated 219 million cases occur in 90 countries. However, at 92% of malaria cases and 93% of malaria deaths, Africa carries a disproportionate ratio of malaria worldwide [1]. However, in recent years, conversations over social media have exploded with online users communicating around varied topics including diseases. The behavioral perspectives of online users and generation of thematic information provides a platform for emergent research surrounding malaria. Increasingly, internet based social media provides a platform for interpersonal and intrapersonal conversations among users who create discourses reflecting current occurrences, for example malaria.

With over 3.5 billion people actively using social media [2], social media plays a significant role in creating novel ways for engaging in discourses, sharing narratives and opinions. When it comes to public health, social media such as Twitter provides avenues to communicate and mine vast amounts of data for purposes of gaining insights on mitigation measures towards emerging epidemic scenarios [3, 4] . Twitter can be described as a microblogging website or service launched in 2006 where users publish messages of up to 140 characters. Millions of tweets (short messages on twitter) are published and read daily on a wide gamut of topics ranging from personal to public thoughts [6, 7]. Further, ubiquitous computing, has encouraged a great deal of interpersonal interaction among community of users over the internet through social media, thereby creating a rich data source for analysis. In fact, statistics indicate that over 90% of data stored today has been generated in the last five years with social media such as YouTube, Facebook, Twitter, and crowdsourcing information contributing a big chunk of the online data [8].

Bureaucracy and lack of adequate resources continue to hinder effective detection and communication of infectious diseases in prone countries [9]. Similarly, the emergence of big data sources such as web 2.0 provide a fertile ground for effective detection and monitoring of infectious diseases [9]. Extracting opinions from data in social media such as Twitter and making meaningful insights for decision making is still a challenge. However, with the advent of mobile telephony and access to affordable Internet access, millions of users can now share and interact with each other irrespective of geographical location. In our study, we demonstrate that Twitter has an enormous potential to gain insights and understanding of public opinion and conversations surrounding malaria.

Our current focus, therefore, is to investigate the behavioral perspectives of online users and emergent conversations around malaria. Online users write short posts about events or current phenomena and engage with each other in near real-time fashion on similar topics of interests thus creating conversations cues and emergent topics which are relevant for topical mining. This study is, therefore, guided by the following research questions (RQ): 1. What are the themes underlying the public perception in terms of sentiments

URI: https://hdl.handle.net/10125/64208
978-0-9981331-3-3

and emotions towards malaria? 2. What are the emergent topical themes and discourses regarding malaria?

2. Literature review

Malaria is an infectious and lethal disease affecting most parts of the world with high prevalent incidences occurring in sub-Saharan Africa with an 81% of worldwide cases and 90% of reported deaths [10]. While there is a limited availability of literature focusing on malaria, existing studies have investigated the impact of infectious diseases such as Zika, and influenza on different geographical populations. For example, latest studies on Zika epidemic have demonstrated the use of Twitter to shed more light on public perceptions concerning Zika with a focus on identifying emergent topics on the negative sentiment category [11]. Miller et al. (2017) conducted an exploratory analysis user tweets along key categories for topical modeling, namely: Treatment, Transmission, Symptoms, and Prevention [27]. Other work focused on infectious diseases such as influenza, Ebola cases in the United states supported by data from Google trends on infectious diseases and statistics from Center for Disease Control and Prevention (CDC) [12]. For example, prior studies on influenza have demonstrated the application of social media, such as, the use of Twitter for detection and surveillance of influenza rates and outbreaks in Canada [13], USA [10, 14], UK [15] and Japan [16]. The researchers used varied methodologies to monitor and detect influenza outbreaks such as Natural Language processing (NLP) specifically sentiment analysis [17], supervised learning using document classification [16], Autoregressive Moving Average (ARMA) model [14], statistical analysis and linear regression [12]. Other studies have used n-gram logistic regression techniques to classify twitter sentiments and Latent Dirichlet Allocation for topical modeling [11].

Previous work in social media literature have investigated public concerns [18], public discourses and discussions [19-21], public health opinions [22] and trends [23, 24]. Recently, the works of Pak et al [7] demonstrated that Twitter data can be an empirical source of information and vehicle to detect and track Influenza occurrences. Similarly, other works on flu trends have proposed near real-time models for detection and prediction of influenza epidemic [10, 12]. Furthermore, studies have also showed that Twitter can be a good source for discovering conversations and trending patterns in the public health domain [25]. Currently, other non-traditional sources of data exist such as dedicated websites and social media platforms such as Facebook, and Twitter where individuals readily share their own experiences of ailments and illnesses before deciding to seek medical services [26].

Application of text mining in social media is considered an important tool in the extraction of timely and efficient information for use by public health practitioners and providers [17]. Consequently, the web is increasingly emerging as a one stop shop for users to search for and share information about infectious diseases [17]. Accordingly, the motivation for this study, therefore, is inspired by the limited availability of literature focusing on infectious diseases such as malaria with higher occurrences afflicting developing nations. In addition, to the best of our knowledge, no studies demonstrate the use of social media as a vehicle to extract discourses surrounding malaria.

3. Research methodology

We used Crimson Hexagon, a social media analytics company to collect relevant tweets. The advantage of using Crimson Hexagon software is that it facilitates data collection and provides insights into online conversations in terms of sentiments and emotions.

3.1. Data collection

The first step is to formulate a query specifying keywords and metadata such as language, source, date range and location etc. A query with a large date range (2008-2019) was developed to assess and extract the relevant topics and conversations emerging from the Twitter corpus. We used a structured query language (SQL) to expound our query to extract relevant data from the Twitter corpus. With an SQL query as shown in Figure 1, we defined the scope for the data collection process by removing retweets, hyperlinks pointing to websites and excluded the occurrence of related diseases such as tuberculosis, cancer, dengue and zika from contaminating the search results as much as possible.

After executing the query, we collected a total of 1,350,732 million relevant Tweets posted only in English language, representing conversations about malaria on Twitter. Our main objective was to answer the research questions RQ 1 and RQ 2 formulated earlier in the introduction section.

```
1  Malaria
2  AND - tuberculosis
3  AND - dengue
4  AND - cancer
5  AND -Zika
6  AND -
7  (RT OR http OR https)
```

Figure 1. Search query

3.2. Data analysis

For data analysis, we used Crimson Hexagon's Crimson Hexagon's ReadME algorithm, a supervised learning method for document classification and opinion mining [28]. The algorithm underlies content analysis and document classification capabilities from unstructured data offered by Crimson Hexagon. The objective is to classify tweets into categories, for example, sentiment analysis focuses on binary impressions of an audience reduced to "positive", "negative" and "neutral" responses. By extension, emotional analysis provides a deeper level of a user's cognitive subjectivities, and impulses represented by a gamut of moods. Sentiment and emotion analysis are two distinct expert methods used to measure the emotional engagement of users. Additionally, Crimson Hexagon allows for powerful exploration of data (slice and dice) with features to analyze trends, and drill down the social media analytics thus answering strategic questions to different audiences.

We first analyzed the sentiments in malaria conversations and classified them as either positive or negative sentiments. Posts lacking identifiable positive or negative sentiments are tagged as neutral. Thus, the neutral sentiments are excluded from the analysis results. Second, we conducted emotion analysis using another classifier included with the Buzz monitor to capture predominant emotions in malaria conversations. The predefined emotions are classified as Sadness, Disgust, Joy, Anger, Fear, and Surprise. Third, we created a new model to include Transmission, Prevention, Symptoms, and Treatment categories. The categories were mainly drawn from existing identifiable disease characteristics commonly used by most medical journals, CDC and WHO [27] to classify different perspectives of infectious diseases. These categories are also supported and reinforced in literature through the works by Miller et al. [27]. We then manually tagged 251 tweets and mapped them across the four categories, following the codebook in Appendix 1. Using Opinion monitor, we then trained the new model to learn and classify the data to reveal perceptions about malaria conversations.

For presentations of the textual data, we used visualizations tools like word clouds and clusters. Word clouds are methods we used in summarizing the frequency of words surrounding malaria conversation on Twitter. Similarly, clusters are visual exploration tools used to summarize the relevant keywords emerging from malaria conversations. The size of the key word shows the relative weight of importance attached in each cluster of words.

4. Results

In this section, the results from the sentiment analysis, emotion analysis and the trained categories are reported. We give explicit analysis and presentation alongside the results guided by our research questions. Also, through topic modeling, results of the emergent topics and themes are explained accordingly.

4.1. Volume and demography of tweets

Overall, with a total of 1,350,732 million English language tweets collected over the period of (May-24-2008 to June-12-2019), Figure 2 gives a summarized graphical representation of the total number of tweets over time. Regarding the demographics of authors who post on Twitter, the identifiable age distribution on the total proportion of tweets collected is shown in Figure 3.

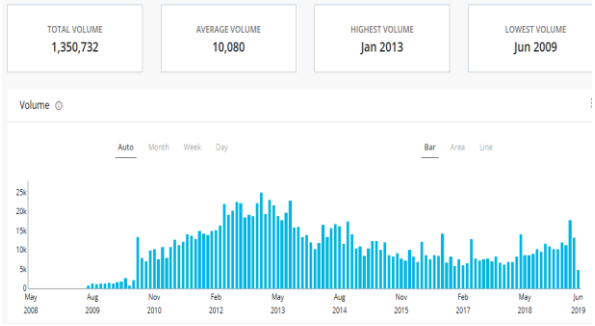

Figure 2. Malaria trend over time

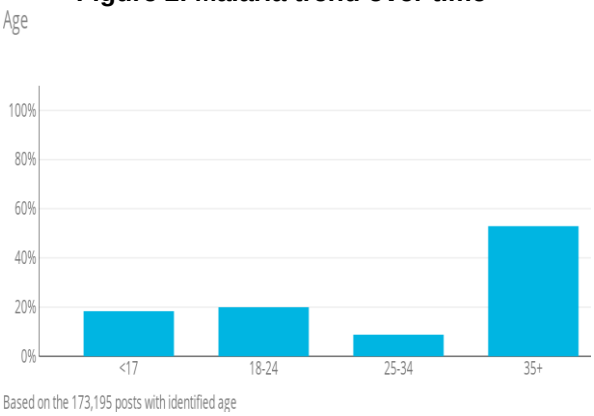

Figure 3. Demographics breakdown.

4.2. Sentiment and emotion analysis

In answer to our first research question concerning the polarity of tweets, 35% of all the tweets showed that the sentiments were negative, 25% were positive, and 40% of the tweets expressed as neutral sentiments as reflected in Figure 4. Additionally, Figure 4 summarizes the volume of tweets spread across the

sentiment categories (positive, negative and neutral). When we exclude 40% of posts that do not contain positive or negative sentiments, the predominant sentiments in malaria conversations become negative as illustrated in Figure 5.

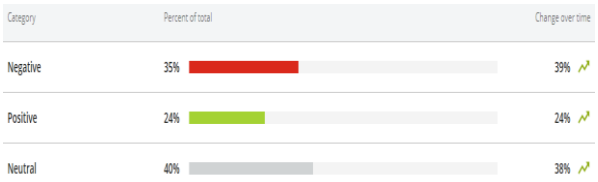

Figure 4. Proportion of sentiments

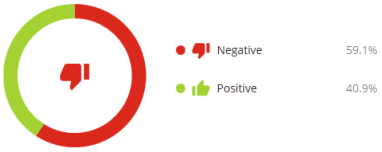

Figure 5. Sentiment wheel

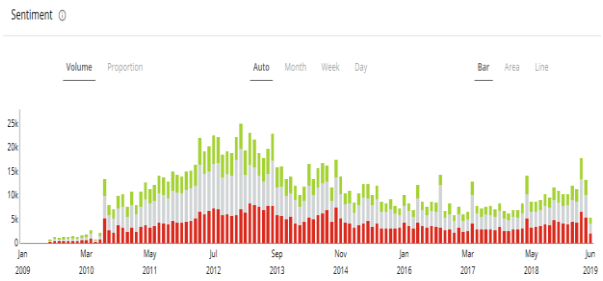

Figure 6. Volume of tweets by years

In addition to sentiment analysis, we used emotion analysis results (Sadness, Disgust, Joy, Anger, Fear and Surprise) to show the proportion of tweets representing the various emotions attached to the discourses surrounding malaria. Figures 7-9. summarizes the results of emotion analysis for all the malaria conversations collected respectively. From Figure 7, the predominant emotions in the malaria conversations include Sadness, Disgust, Joy and Anger. However, 32% of posts that did not contain emotions were excluded. This is also consistent with analysis presented in Figure 9 where the volume of tweets measured over the years are categorized to include Sadness (18%), Disgust (18%), Joy (13%) and Anger (11%).

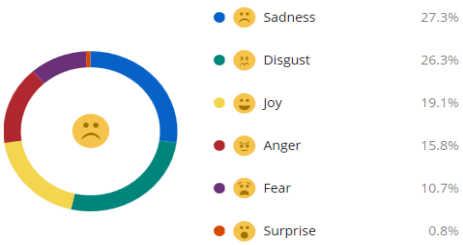

Figure 7. Emotion analysis wheel

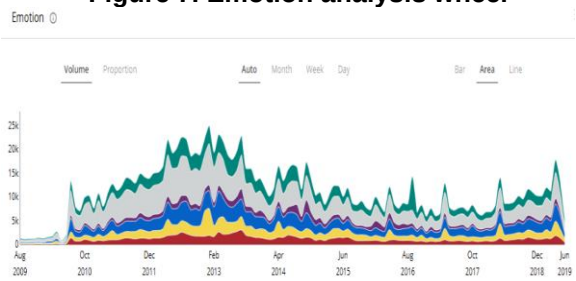

Figure 8. Emotion analysis by volume

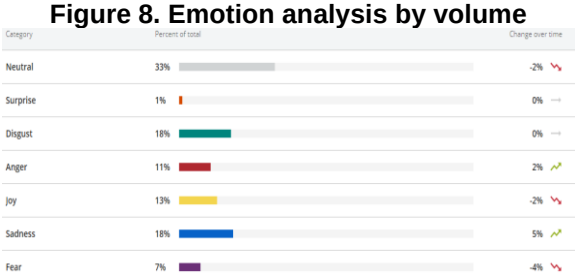

Figure 9. Emotion analysis by percentages

4.3. Geography of Tweets

The collection and analysis of posts included geotagged tweets which provided a rich source of information regarding the location of authors engaged in public discourses concerning malaria. Interestingly, the top two ranks were dominated by public conversation about malaria from USA (33.9%) and UK (16.16%). Other countries in the listings included Nigeria (14.5%), Indonesia (4.94%), India (3.23%) and Kenya (3.15%) as reflected in Table 1.

Table 1. Geographical locations

#	Country	Posts	% of total
1	United States of America	276188	34.06%
2	United Kingdom	129550	15.98%
3	Nigeria	125738	15.51%
4	Indonesia	39099	4.82%
5	India	26190	3.23%
6	Kenya	25526	3.15%
7	Canada	20641	2.55%
8	Ghana	18457	2.28%
9	South Africa	17818	2.20%
10	Australia	15060	1.86%

4.4. Topic modeling

In this section, the sentiment and trained categories are discussed. Next, the emergent topics and themes are developed and summarized in Tables 2-3. respectively. To answer our second research question, we used the crimson hexagon Opinion monitor to define the four categories (Symptoms, Treatment, Prevention and Transmission) and trained them with labelled tweets that we manually identified for the training process. Figure 10 depicts a word cloud representing key topics while Figure 11 depicts a cluster for relevant keywords.

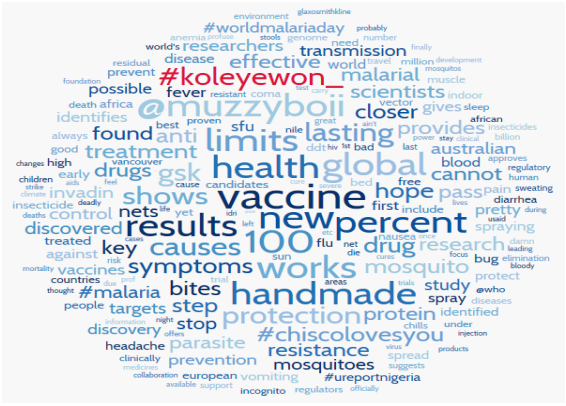

Figure 10. Word cloud showing frequency of words

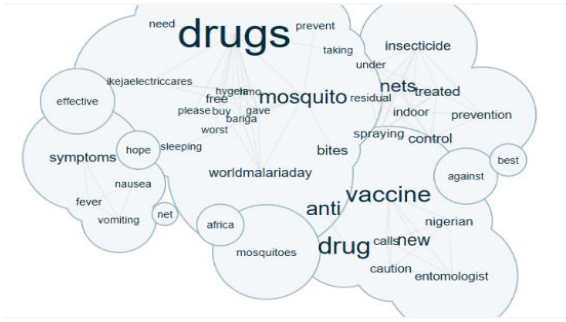

Figure 11. Cluster of keywords for malaria.

4.4.1 Topic modeling by sentiment category. For the sentiment categories, the emergent themes were qualitatively identified into 6 broad categories identified in Table 2. (Awareness, Vaccines, Recommendations, Treatment, Symptoms and Side effects). The formulation of the topics and themes were guided by WHO dictionary on malaria terminology compiled by epidemiology experts with medical knowledge, with a purpose to: 1. Provide adaptable and uniform concepts for use and 2. manage evolution of future concepts and malaria terms [30].

The topics and key words were identified from a combination of clusters and word clouds which represented a collection of posts that were classified as positive. Tweets from the positive sentiment category included words such as “promising”, “achievement”, “great” and “amazing” revealing an optimistic nature of public conversations about ways of combating malaria. The primary themes identified in the positive category reflected public conversations on ways to fight malaria and a discussion on new discoveries and development on malaria vaccines in curbing the spread of the disease.

From the analysis in Figure 5, significant proportions of the tweets were classified as negative sentiments. The emergent topics included: 1. Malaria disease reflecting public conversations regarding the symptoms associated malaria. 2. Malaria medication, which represented topics with public conversations surrounding the effects of taking anti-malaria drugs. Symptoms and side effects were the emergent themes which could provide public health practitioners with a rich source of feedback on self-reported individual opinions. Also, provision of public experiences with malaria symptoms and side effects of anti-malaria medications.

4.4.2 Topic modeling by trained category. The topic modeling results for the four trained categories (Transmission, Prevention, Symptoms, and Treatment) are shown in Table 3. Four emergent themes comprised 1. Malaria infections, 2. Control mechanisms, 3. Body symptoms, and 4. Vaccine development. In the *Transmission* category, *malaria infection* theme was informed by the key topic mosquito bites which reflected several public conversations about being bitten by mosquitos, contracting malaria, and methods of malaria transmission through blood transfusion among others.

The public tweets depicted an engagement among Twitter users about grave concerns on being bitten by mosquitos and a heightened awareness of the different methods of malaria transmission. Regarding, *control mechanisms*, this theme emerged from the trained category *Prevention.* The emergent topic identified was malaria control which reflected the public twitter posts on issues such as use of bed nets to control malaria, employing insecticides and other chemicals to spray mosquito infested areas, dirty environments that could potentially attract breeding of mosquitos among others. Also, use of mobile technologies was reported as well, for example, text messaging as a potential strategy to help control the spread of malaria.

The third trained category, *Symptoms*, yielded malaria signs and symptoms as the emergent topics. Several tweets corroborated this topic to represent public conversations revolving around the different signs and symptoms associated with individuals who reported their experiences and feelings in dealing with

malaria disease. Drilling down deeper into the tweets showed users with emotions ranging from Sadness, Anger and Fear of being infected with malaria.

Lastly, *vaccine development* was identified as the emergent theme in the *Treatment* category. Also, the emergent topic was *malaria drugs* with an inference on public posts generated by user engagement on Twitter on drug related issues. For example, the public tweets reflected on public conversation ranging from new scientific advances in malaria vaccines to how some individuals reported their status as currently taking malaria drugs. Predominant conversation focused mainly on news about current development of affordable drugs or vaccines and mentions of specific drugs, for example *Luter*, to combat malaria.

Overall, the results in Figure 12 shows that the categories symptoms (11.6%), treatment (30.9%), prevention (13.4%) and transmission (24.3%) represent identifiable malaria conversations accounting from the total number of posts extracted from the Twitter corpus.

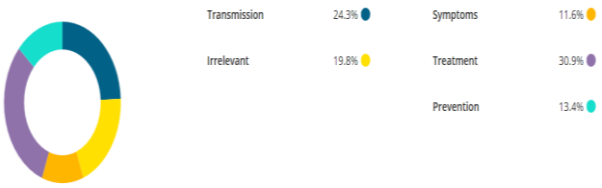

Figure 12. Trained categories

Table 2. Topic modeling for sentiment categories.

Category	Emergent Topics	Key Words	Sample Tweets	Theme
Positive	Fight against malaria	Fight, support, help	"Amazing! Let us know how we can support you as you help fight #malaria!"	**Awareness**
	Malaria tablets	Malaria, tablets, inform,	"Luv friends that inform you right before you fly out of manic episodes Malaria tablets can cause. Love ya"	
	Malaria vaccines	Discover, testing, vaccine, new, trials, scientists	"@ Malaria vaccine to get human trial"	**Vaccines**
			"@ Malaria vaccine great achievement"	
			"@Success in malaria vaccine trials"	
			"@Scientists discover a promising target for a Malaria vaccine"	
Neutral	Recommendations	Guidelines, recommendations, treatment, advice	"Recommendations for malaria treatment are based on specialist guidelines and where appropriate specialist advice for special cases"	**Recommendations**
	Treatment strategies	Strategies, diagnosis, surveillance, control, prevention, detection	"@ Strategies of prevention & control of malaria. 1) surveillance & case management. -Early diagnosis complete treatment -case detection & sentinel surveillance (passive active)"	**Treatment**
			"integrated vector management"	
Negative	Malaria disease	Feverish, terrible, headache, sick, mosquitoes	"Oh no, not this time am feeling feverish since morning n a terrible headache disturbing me. It's been a long time i had malaria!!!"	**Symptoms**
			"Malaria is a really terrible disease to try and fake"	
			"I hate I'm sick. stupid malaria. foolish mosquitoes"	
	Malaria Medication	Hate, worse, taste awful, depressing	"I hate malaria medication. It makes u feel worse than u were"	**Side effects**

Table 3. Topic modeling for trained categories.

Category	Emergent topic	Key words	Sample tweets	Themes
Transmission	Mosquito bites	Positive, contracted, blood transfusion, spread, anopheles	"I have 12 mosquito bites on my body I think I'm going to get malaria I am almost positive!!! Woo!!!"	**Malaria Infections**
			"Pretty sure I have malaria after all these mosquito bites"	
			"Think I have contracted malaria from all those mosquito bites"	
			"Malaria usually spreads via the bite of infected female Anopheles mosquitoes but can also spread via blood transfusions. #WorldMalariaDay"	
Prevention	Malaria control	Control, detect, treat, chemical, messaging, bed nets, environment, spraying, insecticides	"Some ways we can control malaria without DDT are detect malaria early and treat with drugs and use other chemical substitutes"	**Control mechanisms**
			"How text messaging can help control Malaria"	
			"Bed nets are twice as effective at preventing malaria transmission as DDT spraying."	
			"Malaria control. 1. Indoor spraying 2. Insecticide treated bed nets 2. Larviciding 4. Environment measures"	
Symptoms	Malaria signs and symptoms	Headache, nausea, vomiting, severe, muscle pain, diarrhea, signs, cases, death, fever, coma, chills, sweating, fatigue	"Be conversant with the symptoms of malaria which include shaking chills that can range from moderate to severe, high fever, profuse sweating, headache, nausea, vomiting, abdominal pain, diarrhea, anemia, muscle pain, convulsions, coma and bloody stools."	**Body symptoms**
			"Fever, severe cold and headache are few signs of Malaria, just be alert about it. -#MonsoonSmart TRAVEL"	
			"Malaria is curable!!! Know the signs & symptoms. Headache, severe fever and chills, sweating& feeling warm & cold weather, fatigue & dizzy"	
			"In severe cases, malaria can also cause coma and death"	
Treatment	Malaria drugs	New drugs, Luter, treatment, effective, kids, vaccine, advertise, approve, brand, injection, scientists,	"New drug combo treatment for malaria in kids found highly effective"	**Vaccine development**
			"Many anti-malaria drugs target the initial phase of the disease. Luter goes beyond this stage to provide a more robust treatment!! #Luter"	
			"taking my malaria drugs"	
			They advertise new malaria drugs every day. Too many malaria drugs	
			"Medical News Today: European drugs regulators approve world's first malaria vaccine"	
			"Generic anti malaria drugs, hmm no proper brand please"	
			"Malaria treatment begins! Thank God is drugs not injection"	
			"Berlin scientists develop cheap anti-malaria drug"	
			"Fighting malaria drug resistance: Scientists find new way"	

5. Discussion

In this section we discuss the implications of the analysis results of topical modeling for sentiment and trained categories as summarized in Tables 2-3. The discussion is guided and aligned by the two research questions formulated earlier in the introduction section as follows:

5.1. Public perceptions (addressing RQ 1)

The findings of the discovered topics and themes are mapped to the valence of tweets (positive, negative and neutral) summarized in Table 1. For example, within the neutral sentiment category, the themes identified included *recommendations* and *treatment*. The key topics in the neutral sentiment pointed to conversations surrounding recommendation and guidelines alongside treatment strategies for malaria. Evidently, the topics discussed issues concerning how or where to seek treatment advise and strategies to detect, prevent and control malaria. It is unclear why the emergent themes (recommendation and treatment) were categorized as Neutral, perhaps the middle-ground perspective captured user's aspirations to provide lasting solutions to malaria epidemic. In future studies, further refinement of the category can shed more light in terms of tagged tweets and emergent themes.

The geographical distribution of users was important in shaping public perceptions about malaria dialogue. While malaria epidemic occurs mainly in sub-Saharan Africa [1], over 50% of the public conversations on Twitter originated from North America and Europe (see Table 1.). This finding was rather puzzling. Further investigations into the Twitter corpus revealed an interesting story. We found that influencers, donors, organizations, leaders and philanthropists who were based in North America and Europe, contributed actively in public discourse about malaria eradication, for example, 1. dissemination of information; 2. funding various solutions (e.g. resource allocation to eradicate malaria epidemic).

Looking at the volume of tweets over time (see Figure 2), malaria conversations between the year 2011-2013 showed a sharp spike but a gradual decline from 2014-2017. For this, we looked through WHO reports [1] and discovered that various malaria interventions may have provided explanations to this phenomenon. For example, 1. funding declined in 2015-2017 compared to 2012-2014; 2. reduction in malaria mortality rates since 2015 from estimated 607, 000 deaths in 2010 compared to 451,000 in 2016, 3. between 2015-2017, no significant progress was made in reducing malaria cases. From the trend analysis over time (see Figure 1), the malaria conversations may have been attributed to different interventions implemented during each time period.

5.2. Topical themes (addressing RQ 2)

From a health-related point of view, the perceptions derived from the trained categories is consistent with the efforts and strategies currently combated by WHO [1]. When looking at the trained categories (Transmission, Prevention, Symptoms and Treatment), the findings revealed conversations that were consistent with identified disease characteristics in medical journals and by WHO. For example, posts regarding Symptoms correctly captured users who were either affected by malaria or were sharing information about its Symptoms. The implications of this category provide public health officials with valuable insights on behavioral reactions of users to different symptoms. For example, thematic patterns from the Prevention category can improve public health officials' understanding of the effectiveness of different control mechanisms such as application of indoor spraying thus enabling a proactive approach to addressing emerging concerns in a near real-time fashion. Similarly, the insights from Treatment category can provide drug manufacture's with feedback on performance of their products, for example side-effects, and therefore, offer new opportunities in developing effective drug regimens for malaria.

6. Limitations

This study demonstrates that online social media networks, for example, Twitter can be an effective source of rich information for public perceptions (sentiments and emotions) on emergent concepts and themes shaping public discourses on malaria. However, there are several limitations to our study and areas for further improvement. First, the collection and analysis of the Twitter corpus was limited to the English language which has an implication on the generalizability of the study. However, a significant proportion of the total tweets were reported in the English language. Future studies can explore different popular languages like Spanish, Swahili, Malay or Hindi to analyze tweets that may have mentioned malaria and related conversations but were excluded during the data collection process.

Second, while the emergent topics and themes can reflect public discourses on malaria, further text mining techniques, for example NLP can be used to evaluate the Twitter corpus by providing a much more granular control and refinement for additional

improvement of topic and theme development. Lastly, sample bias may be of concern considering the frequency of activities between 'power users' and average users regarding Twitter usage, thus posing a challenge in generalizing results to the general population [31]. However, WHO reports, and statistics can offer a benchmarking guideline to infer on numbers and narratives surrounding malaria.

7. Conclusion

In this study, we address our central research questions by examining the emergent topics, themes and discourses surrounding malaria from Twitter. This study, therefore, is a first effort to shed more light on the collection, analysis and interpretation of conversations and trends around malaria discourse on Twitter. The findings of this study are expected to contribute towards revealing the conversation patterns and trends shaping public opinion regarding malaria dialogue on social media. Additionally, this is one of the first studies to address malaria sentiments, emotions, trends and public discourse analysis using social media, specifically Twitter. The idea of using automated approaches to analyze public sentiments on malaria is to enable public health officials and organizations to distill meaningful insights from public discourse concerning malaria and related issues in a near real-time fashion. For example, the emergent themes and topics from the trained categories can provide insights to inform decisions such as: 1. development of effective drug regimen which could potentially reduce health costs; 2. proactive resource allocation and timely distribution of prevention kits to affected areas; and 3. better or precise management of symptoms and side-effects thus enhancing the overall public healthcare management, which can have implications to other related infectious diseases. To achieve these insights, innovative analytics approaches will help public health professionals answer different questions to include: 1."what happened?"(descriptive analytics), for example, a tweet such as "Fighting malaria drug resistance" and "Scientists find new way" ; 2."What something happened?"(diagnostic analytics); 3."What is likely to happen" (predictive analytics); and 4."what action to take?" (prescriptive analytics). To further improve the analysis, we recommend expanding and refining of keywords for model training to minimize the noise hence enhance the discovery of much more insights from the data. The highlights of the findings will motivate directions for future research, promote large-scale geographical interventions, support public policy decisions and help with development of public health monitors.

8. References

[1] WHO, World Malaria Report. 2018.
[2] Wearsocial, Digital in 2019. 2019.
[3] Cassa, C.A., et al., Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. PLoS currents, 2013. 5.
[4] Chew, C., Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak. 2010.
[5] Lachlan, K.A., P.R. Spence, and X. Lin, Expressions of risk awareness and concern through Twitter: on the utility of using the medium as an indication of audience needs. Computers in Human Behavior, 2014. 35: p. 554-559.
[6] Tumasjan, A., et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. in Fourth international AAAI conference on weblogs and social media. 2010.
[7] Pak, A. and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. in LREc. 2010.
[8] Kass-Hout, T.A. and H. Alhinnawi, Social media in public health. Br Med Bull, 2013. 108(1): p. 5-24.
[9] Littrell, M., et al., Documenting malaria case management coverage in Zambia: a systems effectiveness approach. Malaria journal, 2013. 12(1): p. 371.
[10] Achrekar, H., et al. Predicting flu trends using twitter data. in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. 2011. IEEE.
[11] Mamidi, R., et al., A Study to Identify Key Topics Bearing Negative Sentiment on Twitter Concerning the 2015/2016 Zika Epidemic.
[12] Chen, L., et al. Vision: towards real time epidemic vigilance through online social networks: introducing SNEFT--social network enabled flu trends. in Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond. 2010. ACM.
[13] Byrd, K., A. Mansurov, and O. Baysal. Mining twitter data for influenza detection and surveillance. in Proceedings of the International Workshop on Software Engineering in Healthcare Systems. 2016. ACM.
[14] Culotta, A. Towards detecting influenza epidemics by analyzing Twitter messages. in Proceedings of the first workshop on social media analytics. 2010. ACM.
[15] Lampos, V. and N. Cristianini. Tracking the flu pandemic by monitoring the social web. in Cognitive Information Processing (CIP), 2010 2nd International Workshop on. 2010. IEEE.
[16] Aramaki, E., S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. in Proceedings of the conference on empirical methods in natural language processing. 2011. Association for Computational Linguistics.
[17] Gomide, J., et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. in Proceedings of the 3rd international web science conference. 2011. ACM.
[18] Lazard, A.J., et al., Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. American journal of infection control, 2015. 43(10): p. 1109-1111.

[19] Stefanidis, A., et al., Zika in Twitter: temporal variations of locations, actors, and concepts. JMIR public health and surveillance, 2017. 3(2).
[20] Orr, D., A. Baram-Tsabari, and K. Landsman, Social media as a platform for health-related public debates and discussions: the Polio vaccine on Facebook. Israel journal of health policy research, 2016. 5(1): p. 34.
[21] Lazard, A.J., et al., Public reactions to e-cigarette regulations on Twitter: a text mining analysis. Tobacco control, 2017. 26(e2): p. e112-e116.
[22] Karami, A., et al., Characterizing diabetes, diet, exercise, and obesity comments on Twitter. International Journal of Information Management, 2018. 38(1): p. 1-6.
[23] Khatua, A. and A. Khatua. Immediate and long-term effects of 2016 Zika outbreak: a twitter-based study. in 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom). 2016. IEEE.
[24] Alicino, C., et al., Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infectious diseases of poverty, 2015. 4(1): p. 54.
[25] Prier, K.W., et al. Identifying health-related topics on twitter. in International conference on social computing, behavioral-cultural modeling, and prediction. 2011. Springer.
[26] Sharpe, J.D., et al., Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR public health and surveillance, 2016. 2(2).
[27] Miller, M., et al., What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR public health and surveillance, 2017. 3(2): p. e38.
[28] Hopkins, D.J. and G. King, A method of automated nonparametric content analysis for social science. American Journal of Political Science, 2010. 54(1): p. 229-247.
[29] Hearst, M.A. Untangling text data mining. in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999. Association for Computational Linguistics.
[30] Organization, W.H., WHO malaria terminology. 2016.
[31] Bian, J., et al., Mining twitter to Assess the public perception of the "Internet of Things". PloS one, 2016. 11(7): p. e0158450.
[32] Bush, A.O., et al., Parasitism: the diversity and ecology of animal parasites. 2001: Cambridge University Press.
[33] NCI Dictionary, "Definition of symptom - NCI Dictionary of Cancer Terms - National Cancer Institute." [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/symptom. [Accessed: 06-Sep-2019].
[34] Merriam-Webster, "Treatment | Definition of Treatment by Merriam-Webster." [Online]. Available: https://www.merriam-webster.com/dictionary/treatment. [Accessed: 06-Sep-2019].

Appendix 1: Codebook for labeling topic modeling by trained category

Category	Description	Keywords	Examples
Transmission	Refers to the passing of a pathogen causing communicable disease from an infected host individual or group to an individual or group, regardless of whether the other individual was previously infected [32].	Blood transfusion, Spread anopheles	"All these bug bites. I think I have malaria." "Pretty sure I have malaria from all the mosquitos that bit me tonight"
Prevention	Refers to the action taken to decrease the chance of getting a disease or condition [33].	Control, chemical, bed nets, spraying, insecticides	"incognito is the ONLY spray that is clinically proven to protect against malaria" "Sleeping under long-lasting insecticidal nets protects against malaria"
Symptoms	A physical or mental problem that a person experiences that may indicate a disease or condition. Some examples of symptoms are headache, fatigue, nausea, and pain [33].	Headache, nausea, vomiting, fatigue, pain, death, fever.	"#Malaria101 (6) Malaria symptoms include fever, chills, headaches, muscle aches, nausea and vomiting @fightingmalaria"
Treatment	Refers to a substance or technique used in treating [34].	drugs, treatment, vaccine, brand, injection, scientists,	"Malaria symptoms!! taking injection on doc prescription*crying*"