Classifying Information from Microblogs during Epidemics

Koustav Rudra
IIT Kharagpur, India
koustav.rudra@cse.iitkgp.ernet.in

Ashish Sharma
IIT Kharagpur, India
ashishshrma22@gmail.com

Niloy Ganguly
IIT Kharagpur, India
niloy@cse.iitkgp.ernet.in

Muhammad Imran
Qatar Computing Research Institute, HBKU, Doha, Qatar
mimran@hbku.edu.qa

ABSTRACT

At the outbreak of an epidemic, affected communities want/need to get aware of disease symptoms, preventive measures, and treatment strategies. On the other hand, health organizations try to get situational updates to assess the severity of the outbreak, known affected cases, and other details. Recent emergence of social media platforms such as Twitter provide convenient ways and fast access to disseminate and consume information to/from a wider audience. Research studies have shown potential of this online information to address information needs of concerned authorities during outbreaks, epidemics, and pandemics. In this work, we target three communities (i) people who are not affected yet and are looking for prevention-related information (ii) people who are affected and looking for treatment-related information, and (iii) health organizations like WHO, who are interested in gaining situational awareness to make timely decisions. We use Twitter data from two recent outbreaks (Ebola and MERS) to built an automatic classification approach using low level lexical features which are useful to categorize tweets into different disease-related categories.

1 INTRODUCTION

With the wide-adaptation of Information Communication Technology (ICT) and social media platforms, information dissemination and access to up-to-date information become easier. General public post textual messages on social networks to report what they observe and hear around them. During disease outbreaks, information posted on microblogging platforms such as Twitter by affected communities provide rapid access to diverse and useful insights helpful to understand various aspects of the outbreak. Research studies conducted with formal health organizations have shown potential of such health-related information on Twitter for quick response [3].

However, in order to effectively use this online information for any type of response efforts or decision-making processes, during an ongoing epidemic situation, it is essential to process and analyze tweets as they arrive (i.e., in near real-time). In an epidemic, various types of information, including disease-related updates and personal opinion about the disease are posted by users in huge volume and at rapid rates. This online content contains valuable but multi-dimensional information like ‘disease sign and symptoms’, ‘prevention mechanism’, ‘transmission mediums’‘death reports’ etc. To make it presentable to health experts, these tweets must be automatically classified into some informative categories (e.g. symptom reports, prevention, treatment, etc).

Moreover, we observe that different stakeholders (e.g. different health organizations and affected or vulnerable communities) have different information needs and the above mentioned categories can satisfy the requirement of different stakeholders. In this paper, we target the following three communities —

(1) **Pre-disease community:** people who are primarily looking for preventive measures, signs or symptoms of a disease to take precautionary measures. These are not affected people but they are vulnerable.
(2) **Post-disease community:** In this case, the target community is considered already affected by the epidemic (e.g. users have already fallen sick). The users in this community look for treatment-related information or find nearby hospitals which deal with the issues of concern.
(3) **Monitoring Organizations:** During epidemics, government and other health monitoring agencies (WHO, CDC) look for information about victims, affected people or death reports, vulnerable people etc. This information is used to determine the severity of the situation and accordingly necessary actions such as employing experts/doctors from other countries, setting up separate treatment centers, are planned.

To the best of our knowledge, all previous research works that utilize social media for health [3, 22, 26] focus on analyzing behavioral and social aspects of users who post information about a particular disease and to predict whether a user is likely to catch the disease in future based on their posts. However, we believe that social media contains more important information that can be used to assist different communities (e.g., affected or vulnerable population, health organizations) for a number of purposes (e.g., extraction of signs and symptoms, treatment advises), if processed timely and effectively.

For this purpose, one potential approach is to perform automatic classification of the messages using supervised machine learning techniques. However, training supervised models require human-annotated data and understanding of the characteristics of messages that distinguish them from one category to another. To this end,

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07...$15.00
https://doi.org/10.1145/3079452.3079491

we first aim to understand the low-level lexical features which can be used to distinguish between different disease categories and stages (i.e. pre, post). Based on the identified features, we employ supervised machine learning technique to develop a classifier. However, building machine learning classifier requires human-labeled examples, which are often not available at the outbreak of a disease. In order to address the issue of labeled example scarcity, another challenge that we tackle in this work is to develop generic classifier applicable for multiple diseases. To this end, our vocabulary-independent features allow the classifier to perform accurately in cross-domain scenarios, e.g., when the classifier trained over tweets posted during a past outbreak is used to predict tweets of a future/current outbreak.

To evaluate, we check the performance of our proposed classifier using real-world Twitter datasets collected during two recent disease outbreaks namely Ebola and Middle East Respiratory Syndrome (MERS) [19, 20]. Low-level lexical features perform quite well compared to vocabulary based approach [9] specially in cross-domain scenario (section 4).

The rest of the paper is organized as follows. In the next section, we summarize related work in this domain. Section 3 provides details regarding our datasets and classification categories. Section 4 describes human-annotation and feature learning steps along with the results. We conclude the paper in section 5.

2 RELATED WORK

Twitter, Facebook, online health forums and message boards are increasingly being used by professionals and patients to obtain health information and share their health experiences [11]. Fox et al [21] reported that 80% of the internet users have utilized online available information on health-related topics like disease symptoms, diagnosis or treatment. The popularity of social media in medical and health domain has gained attention from researchers for studying various topics on healthcare. In this section, we provide a brief overview of them.

Various methods have been proposed for mining health and medical information from clinical notes. Most of these works have focused on extracting a broad class of medical conditions (e.g., diseases, injuries, and medical symptoms) and responses (e.g., diagnoses, procedures, and drugs), with the goal of developing applications that improve patient care [4, 8, 24]. Recently, Goodwin et al [6] proposed a clinical question-answering system.

Scanfield et al [17] used Q-Methodology to determine the main categories of content contained in Twitter users' status updates mentioning antibiotics. Lu et al [12] built a framework based on clustering analysis technique to explore interesting health-related topics from online health community. Recently large scale researches have been done in exploring how microblogs can be used to extract symptoms related to disease [13], mental health [7] and so on. Most of the methods proposed for extracting information from clinical text utilize earlier proposed systems (e.g. MetaMap [1], cTakes [16]) for mapping clinical documents to concepts of medical terminologies and ontologies (e.g. UMLS [23], SNOMED CT [25]). However, tools like MetaMap were designed specifically to process clinical documents and user-generated medical text from social media differ significantly from formal texts. Recent studies have shown that directly applying Metamap on Social Media Data leads to low quality word labels [18].

Now a days, microblogs provide vast range of real time information about current events, disasters, epidemics etc. These microblogs are not written in a formal way; rather they contain spelling mistakes, noises, many shorthand expressions etc. Researchers put lot of effort in designing text classification techniques [9, 15] suitable for microblogs. To our knowledge, all the existing methods try to extract knowledge from past medical records. In this work, we try to propose a realtime classifier which can classify short microblog messages posted during epidemics into various informative categories.

3 DATASET AND CLASSIFICATION OF MESSAGES

This section describes the datasets of tweets that are used to evaluate our classification approach.

3.1 Epidemics

We collected the crisis-related messages using AIDR platform [9] from Twitter posted during two recent epidemics.

(1) **Ebola:** This dataset consists of 5.08 million messages posted between August 6th, 2014 and January 19th, 2015 obtained using different keywords (e.g., #Ebola, $\cdots$).
(2) **MERS:** This dataset was collected during Middle East Respiratory Syndrome (MERS) outbreak, which consists of 0.215 million messages posted between April 27th and July 16th, 2014 obtained using different keywords (e.g., #MERS, $\cdots$)

For each event, we selected the first 200,000 English tweets in chronological order. The language of a tweet is checked using the language information provided by Twitter.

3.2 Types of tweets posted during epidemics

As stated earlier, tweets posted during an epidemic event include disease-related tweets as well as non-disease tweets. We employed human volunteers to observe different categories of the tweets and asked them to annotate accordingly (details in Section 4). The different disease categories our volunteers identified (which agrees with prior works [6, 10]) are as follows. Some example tweets of each category are shown in Table 1.

Disease-related Messages: Messages which contain disease-related information are primarily of the following five types:

(1) *Symptom* – reports of symptoms such as fever, cough, diarrhea, and shortness of breath or questions related to these symptoms.
(2) *Prevention* – questions or suggestions related to the prevention of disease or mention of a new prevention strategy.
(3) *Disease transmission* – reports of disease transmission or questions related to disease transmission.
(4) *Treatment* – questions or suggestions regarding the treatments of the disease.
(5) *Death report* – reports of affected people due to the disease.

Table 1: Examples of various types of disease tweets (which contribute to information about epidemic) and non-disease tweets.

Type	Event	Tweet text
Disease tweets (which contribute to information about epidemic)		
Symptom	Ebola	early #ebola symptoms include fever headache body aches cough stomach pain vomiting and diarrhea
	MERS	middle east respiratory syndrome symptoms include cough fever can lead to pneumonia & kidney failure
Prevention	Ebola	ebola is a deadly disease prevent it today drink / bath with salty warm water
	MERS	#mers prevention tip 3/5 | avoid touching your eyes nose and mouth with unwashed hands
Disease transmission	Ebola	airborne cdc now confirms concerns of airborne transmission of ebola
	MERS	world health a camel reasons corona virus transmission
Treatment	Ebola	tygerberg hospital is ready 2 treat ebola disease
	MERS	cn-old drugs tested to fight new disease mers
Death report	Ebola	the largest #ebola outbreak on record has killed 4,000+
	MERS	saudia arabia reports 102 deaths from mers disease
Non-disease tweets		
Not relevant	Ebola	lies then he came to attack nigeria with ebola disease what is govt doing about that too
	MERS	good question unfortunately i have not the answer but something to investigate fomites #mers

Table 2: Number of tweets present in different classes.

Event	Symptom	Prevention	Transmission	Treatment	Death report	Non disease
Ebola	52	69	65	59	51	56
MERS	105	70	77	74	68	84

Non-disease Messages: Tweets which do not contribute to disease awareness, most of the time containing sentiment/opinion of common people.

In this work, we try to classify tweets into various informative categories as stated above. We describe our classification technique in the next section.

4 CLASSIFICATION OF TWEETS

As stated earlier, in this section we try to classify tweets posted during epidemic into following classes — (i). symptom, (ii). prevention, (iii) transmission, (iv) treatment, (v) death report, and (vi) non-disease. We follow a supervised classification approach for which we need a gold standard of labeled tweets.

4.1 Gold Standard

For training the classifier, we considered 2000 randomly selected tweets (after removing duplicates and retweets) related to both the events. Three human volunteers independently observed the tweets, deciding whether they contribute to information about epidemic.[1] We obtained unanimous agreement (i.e., all three volunteers labeled a tweet similarly) for 87% of the tweets. For rest of the tweets, we follow majority verdict. Non-disease category contain large number of tweets compared to other classes. Hence, we discard the large number of extra tweets present in *non-disease* for tackling class imbalance. Table 2 shows the number of tweets in the gold standard created.

[1] All volunteers are regular users of Twitter, have a good knowledge of the English language.

4.2 Classification features

We aim to build a classifier which can be trained over tweets posted during past disease outbreaks and then directly can be used over tweets posted for future epidemics. Earlier Rudra et al. [15] showed that low level lexical features are useful in developing event independent classifier and they can outperform vocabulary based approaches. Hence, we take the approach of using a set of event independent lexical and syntactic features for the classification task.

A disease independent classification of tweets requires lexical resources which provide domain knowledge and associated terms. In this work, we have considered large medical knowledgebase Unified Medical Language System (UMLS) [23] comprises over 3 million concepts, each of which is assigned to atleast one of the 134 semantic type. Next, MetaMap [1] is used for mapping texts to UMLS concepts. We perform pre-processing of the data by removing unnecessary words (URLs, mentions, hashtag signs, emoticons, punctuation, and other Twitter specific tags) from the tweets. For this purpose, we used the Twitter POS tagger [5]. After preprocessing tweets are passed as input to MetaMap which returns the set of tokens present in the tweet as concepts in UMLS Metathesaurus along with their corresponding semantic type. As mentioned in section 2, Metamap does not perform well in case of short, informal texts. Hence, for this study, we only consider those words which are formal English words and present in an English dictionary [2]. Finally, semantic types obtained from Metamap are utilized for finding the relevant features. Table 3 lists the classification features (binary).

4.3 Performance

We compare performance of our proposed set of lexical features with a standard Bag-of-Words (BOW) model similar to that in [9] where unigrams are considered as features. We have removed (URLs, mentions, hashtag signs, emoticons, punctuation, stopwords, and other Twitter specific tags) from the tweets using Twitter pos tagger [5].

Model selection: For this experiment, we consider four state-of-the-art classification models from Scikit-learn package [14] —

Table 3: Lexical features used to classify tweets across different classes.

Feature	Explanation
Presence of sign/symptoms	We check if a concept ('phsf', 'sosy') related to symptoms is present in the tweet. Expected to be higher in symptom related tweets. The semantic types which indicate the presence of such term are Sign or Symptom; Physiologic Function
Presence of preventive procedures	Concepts related to preventive procedures ('topp') mostly present in preventive category tweets.
Presence of anatomy	preventive procedures sometimes indicate taking care of certain parts of body. This feature identifies the presence of terms related to body system, substance, junction, body part, organ, or organ Component. Concepts like 'bdsu', 'blor', 'bpoc' are present in tweets describing anatomical structures.
Presence of preventive terms	Terms like 'preventive', 'prevention' etc. indicates tweets containing information about preventive mechanism.
Presence of transmission terms	Terms like 'transmission', 'spread' mostly present in tweets related to disease transmission.
Presence of treatment terms	Terms like 'treating', 'treatment' mostly present in tweets related to treatment.
Presence of death terms	Tweets related to dead people contains terms like 'die', 'kill', 'death' etc.

Table 4: Classification accuracies of tweets, using (i) bag-of-words features (BOW), (ii) proposed features (PRO). Diagonal entries are for in-domain classification, while the non-diagonal entries are for cross-domain classification.

Train set	Test set			
	Ebola		MERS	
	BOW	PRO	BOW	PRO
Ebola	**84.78%**	84.02%	65.69%	**76.15%**
MERS	66.19%	**74.72%**	**88.26%**	81.05%

(i). Support Vector Machine (SVM) classifier with the default RBF kernel and gamma = 0.5, (ii). SVM classifier with linear kernel and $l2$ optimizer, (iii). Logistic regression, and (iv). Naive-Bayes classifier. SVM classifier with RBF kernel outperforms other classification models in case of proposed set of features and Logistic regression model shows best performance where unigrams are considered as features. Hence, we take following two classification models for rest of the study.

We compare the performance of the two feature-sets under two different scenarios (i) in-domain classification, where the tweets of same disease are used to train and test the classifier using 10-fold cross validation, and (ii) cross-domain classification, where the classifier is trained with tweets of one disease, and tested on another disease. Table 4 shows the accuracies of the classifier using bag-of-words model (BOW) and the proposed features (PRO) on the tweets.

In-domain classification: BOW model performs well in the case of in-domain classification (diagonal entries in Table 4) due to uniform vocabulary used during a particular event. However, performance of the proposed lexical features is at par with the bag-of-words model.

Cross-domain classification: The non-diagonal entries of Table 4 represent the accuracies, where the event stated on the left-hand side of the table represents the training event, and the event stated at the top represents the test event. The proposed model performs better than the BOW model in such scenarios, since it is independent of the vocabulary of specific events. For cross-domain classification, we have also measured precision, recall, f-score of classification for both set of features. Due to class imbalance problem, we consider weighted measure for precision, recall, and f-score which takes care of class imbalance. Table 5 shows recall, and f-score for each set of features where left hand side represents training event and right hand side represents test event. Our proposed set of features achieve high recall and f-score compared to bag-of-words model which indicates low level lexical features can show promising performance in classifying tweets posted during future epidemics.

Table 5: Recall (F-score) of tweets, using (i) bag-of-words features (BOW), (ii) proposed features (PRO).

Train set	Test set			
	Ebola		MERS	
	BOW	PRO	BOW	PRO
Ebola	0.84(0.85)	0.84(0.84)	0.65(0.66)	0.76(0.76)
MERS	0.66(0.65)	0.75(0.75)	0.88(0.88)	0.81(0.81)

4.4 Analyzing misclassified tweets

From table 4, it is clear that in cross-domain scenario around 25% tweets are misclassified. In this part, we analyze different kind of errors present in the data and also identify the reasons behind such misclassification. It is observed that in most of the cases tweets from symptom, prevention, and transmission classes are wrongly tagged as not relevant due to absence of the features presented in table 3. When we train our proposed model using Ebola data and test it over MERS, tweets belonging to symptom, prevention, and disease transmission classes are misclassified as non-disease in 12%, 13% and 8% of the cases respectively. Sometimes, same tweet contains information about more than one class like both symptoms and prevention or symptoms and transmission. In such cases, classifier is confused and selects a label arbitrarily. Table 6 shows examples of misclassified tweets, with their true and predicted labels. In most of the cases, we need some features which can discriminate between two closely related classes. In future, we will try to incorporate more low level lexical features to improve classification accuracy.

Table 6: Examples of misclassified tweets.

Tweet	True class	Predicted class
worried about the #mers #virus here are 10 ways to boost your body's immune system to fight disease #health	prevention	not relevant
the truth is that #coronavirus #mers can transmit between humans we think not as well as flu but protect yourself anyway wash hands 24/7	prevention	disease transmission
from on mers-cov wash your hands cover your coughs and sneezes and stay home if you are sick	prevention	symptom
learn more about #mers the virus that causes it how it spreads symptoms prevention tips & amp what cdc is doing	symptom	prevention
wash your hands folks and keep your areas clean mers-middle east respiratory syndrome 1/3 of the people who get this dies	prevention	death reports
#mers is not as contagious as the flu says #infectiousdisease expert via	disease transmission	not relevant

5 CONCLUSION

Sudden disease outbreaks bring challenges for vulnerable and affected communities to find answers to their rapid questions like what are the symptoms of the disease, preventive measures, and treatment strategies. Health organizations also look for situational updates from affected population to prepare response. In this work, We target three communities; vulnerable people, affected people, and health organizations. To provide precise and timely information to these communities, we have presented a classification approach to extract useful information from a microblogging platform during outbreaks. The proposed classification approach uses low-level lexical class-specific features to effectively categorize raw Twitter messages. We developed a domain-independent classifier which performs better than domain-dependent bag-of-words technique specially in cross-domain scenario. Extensive experimentation conducted on real-world Twitter datasets from Ebola and MERS outbreaks show the effectiveness of the proposed approach. The number of messages classified in each category are still quite large and beyond the scope of human processing. In future, we will try to summarize those classified information so that different stakeholders can quickly get an overview of the current situation.

6 COMPETING INTERESTS

The authors don't have any competing interests in this paper.

Acknowledgement: K. Rudra was supported by a fellowship from Tata Consultancy Services.

REFERENCES

[1] A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.

[2] aspell 2015. Aspell-python — Python wrapper for aspell (C extension and python version). (2015). https://github.com/WojciechMula/aspell-python.

[3] M. De. Choudhury. 2015. Anorexia on tumblr: A characterization study. In *Proc. ACM Digital Health*. ACM, 43–50.

[4] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. 2004. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association* 11, 5 (2004), 392–402.

[5] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc. ACL*. Association for Computational Linguistics, 42–47.

[6] T. R. Goodwin and S. M. Harabagiu. 2016. Medical Question Answering for Clinical Decision Support. In *Proc. CIKM*. ACM, 297–306.

[7] C. M. Homan, N. Lu, X. Tu, M. C. Lytle, and V. Silenzio. 2014. Social structure and depression in TrevorSpace. In *Proc. CSCW*. ACM, 615–625.

[8] G. Hripcsak, J. H. Austin, P. O. Alderson, and C. Friedman. 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. *Radiology* 224, 1 (2002), 157–163.

[9] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proc. WWW companion*. 159–162.

[10] M. Imran, P. Mitra, and C. Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proc. LREC*.

[11] N. A. Kinnane and D. J. Milne. 2010. The role of the Internet in supporting and informing carers of people with cancer: a literature review. *Supportive Care in Cancer* 18, 9 (2010), 1123–1136.

[12] Y. Lu, P. Zhang, and S. Deng. 2013. Exploring health-related topics in online health community using cluster analysis. In *System Sciences (HICSS), 2013 46th Hawaii International Conference*. IEEE, 802–811.

[13] M. J. Paul and M. Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health.. In *Proc. Icwsm*, Vol. 20. 265–272.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[15] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh. 2015. Extracting Situational Information from Microblogs during Disaster Events: a Classification-Summarization Approach. In *Proc. CIKM*.

[16] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513.

[17] D. Scanfeld, V. Scanfeld, and E. L. Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control* 38, 3 (2010), 182–188.

[18] H. Tu, Z. Ma, A. Sun, and X. Wang. 2016. When MetaMap Meets Social Media in Healthcare: Are the Word Labels Correct? In *Information Retrieval Technology*. Springer, 356–362.

[19] mers 2014. Centers for Disease Control and Prevention. (2014). https://www.cdc.gov/coronavirus/mers/.

[20] ebola 2014. World Health Organization (WHO). (2014). http://www.who.int/mediacentre/factsheets/fs103/en/.

[21] Susannah Fox. 2011. *The social life of health information, 2011.* Pew Internet & American Life Project Washington, DC.

[22] Ed de Quincey, Theocharis Kyriacou, and Thomas Pantin. 2016. # hayfever; A Longitudinal Study into Hay Fever Related Tweets in the UK. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 85–89.

[23] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl 1 (2004), D267–D270.

[24] Daniel T Heinze, Mark L Morsch, and John Holbrook. 2001. Mining free-text medical records.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 254.

[25] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. SNOMED clinical terms: overview of the development process and project status.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 662.

[26] E. Yom-Tov. 2015. Ebola data from the Internet: An opportunity for syndromic surveillance or a news event?. In *Proc. ACM Digital Health*. ACM, 115–119.