Information Systems Frontiers (2018) 20:933–948
https://doi.org/10.1007/s10796-018-9844-9

Classifying and Summarizing Information from Microblogs During Epidemics

Koustav Rudra[1] · Ashish Sharma[1] · Niloy Ganguly[1] · Muhammad Imran[2]

Published online: 20 March 2018

Abstract

During a new disease outbreak, frustration and uncertainties among affected and vulnerable population increase. Affected communities look for known symptoms, prevention 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 types of end-users (i) vulnerable population—people who are not yet affected and are looking for prevention related information (ii) affected population—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 build an automatic classification approach useful to categorize tweets into different disease related categories. Moreover, the classified messages are used to generate different kinds of summaries useful for affected and vulnerable communities as well as health organizations. Results obtained from extensive experimentation show the effectiveness of the proposed approach.

Keywords Health crisis · Epidemic · Twitter · Classification · Summarization

1 Introduction

During disease outbreaks, information posted on microblogging platforms such as Twitter by affected people provide rapid access to diverse and useful insights helpful to understand various facets of the outbreak. Research studies conducted with formal health organizations have shown the potential of such health-related information on Twitter for a quick response (De Choudhury 2015). However, during an ongoing epidemic situation, in order for health organizations to effectively use this online information for response efforts or decision-making processes, the information should be processed and analyzed as quickly as possible. During an epidemic, social media users post millions of messages containing information about disease sign and symptoms, prevention strategies, transmission mediums, death reports, personal opinions and experiences.

To enable health experts understand and use this online information for decision making, messages must be categorized into different informative categories (e.g. symptom reports, prevention reports, treatment reports) and irrelevant content should be discarded. Although the categorization step helps organize related messages into categories, each category may still contain thousands of messages which would again be difficult to manually process by health experts

This work is an extended version of the short paper: Rudra et al., Classifying Information from Microblogs during Epidemics, Proceedings of the 7th ACM International Conference on Digital Health, 2017.

✉ Koustav Rudra
koustav.rudra@cse.iitkgp.ernet.in

Ashish Sharma
ashishshrma22@gmail.com

Niloy Ganguly
niloy@cse.iitkgp.ernet.in

Muhammad Imran
mimran@hbku.edu.qa

[1] IIT Kharagpur, Kharagpur, India

[2] Qatar Computing Research Institute, HBKU, Doha, Qatar

as well as by affected or vulnerable people. While the key information contained in these tweets is useful for the health experts in their decision-making process, we also observe that different disease categories contain different traits (e.g., specific symptoms characteristics), which can be exploited in order to extract and summarize relevant information.

Moreover, we observe that different stakeholders (e.g. health organizations and affected or vulnerable user groups) have different information needs. In this paper, we target the following three user groups/population—(i) **Vulnerable population:** 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. (ii) **Affected population:** people who are already under the influence of disease and trying to recover from the situation. (iii) **Health organizations:** primarily government and health organizations who look for general situational updates like 'how many people died or under treatment', 'any new service required', etc.

Assisting Vulnerable Population The vulnerable groups look out for precautionary measures which can guard them against acquiring a disease. Our proposed system tries to extract various small-scale precautionary measures like signs and symptoms (such as 'fever', 'flu', 'vomiting', 'diarrhea'), disease transmission mediums ('Transmission of Ebola Virus By Air Possible') etc., from related informational categories of tweets in order to assist these vulnerable groups. Automatic extraction of such precautionary measures or symptoms is a challenging task due to a number of reasons. For instance, we observe that such an information is floated in two flavors—(i) positive (confirmations): e.g. someone confirms that "flu" is a symptom of the Ebola virus or a tweet reports that people should follow x,y,z measures to avoid getting affected (ii) negative (contradictions): e.g. a tweet reports that "fever" is not a symptom of the Ebola virus. In this case, our system should clearly specify that "fever" is not a symptom of Ebola. In order to effectively tackle this, our system extracts the contextual information (positive or negative) of the terms related to precautionary measures such as symptoms, preventive suggestions and accordingly assists people during epidemics.

Assisting Affected Population 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 ongoing epidemic. In order to assist these users, we extract recovery and treatment information from tweets. In case of contagious diseases, it is necessary to alert the affected user groups so that further transmission of the disease can be stopped.

Assisting Health Organizations During epidemics, government and other health monitoring agencies (WHO, CDC) look for information about victims, affected people, death reports, vulnerable people etc. so that they can judge the severity of the situation and accordingly take necessary actions like taking help of experts/doctors from other countries, setting up separate treatment centers. Many a time, travelers from foreign countries also get affected by sudden outbreaks. In such cases, local government has to inform their respective countries about the current status; sometimes they have to arrange special flights to send the affected people to their home countries. Considering all these perspectives, the proposed approach tries to extract relevant information about affected or vulnerable people.

To the best of our knowledge, all previous research works regarding health and social media (De Choudhury 2015; de Quincey et al. 2016; Yom-Tov 2015) focus on analyzing behavioral and social aspects of users who post information about a particular disease and predict whether a user is going to encounter such disease in future based on her current posts. However, a generic model which could assist different stakeholders during an epidemic is important. We make the following contributions in this work:

Contributions

- We develop a classifier which uses low-level lexical features to distinguish between different disease categories. Vocabulary independent features allow our classifier to function accurately in cross-domain scenarios, e.g., when the classifier trained over tweets posted during some past outbreak is used to predict tweets posted during a future/current outbreak.
- From each of the identified information classes, we propose different information extraction-summarization techniques, which optimize the coverage of specific disease related terms using an Integer Linear Programming (ILP) approach. Information extracted in this phase helps fulfill information needs of different affected or vulnerable end-users.

Note that, our epidemic tweet classification approach was first proposed in a prior study (Rudra et al. 2017). The present work extends our prior work as follows. After classification of tweets into different informative categories (symptom, prevention etc.), we propose novel ILP based information extraction-summarization methods for each of the classes which extracts disease related terms and maximizes their coverage in final summary.

Figure 1 provides an overview of our approach. Experiments conducted over real Twitter datasets on two recent disease outbreaks (World Health Organization (WHO)

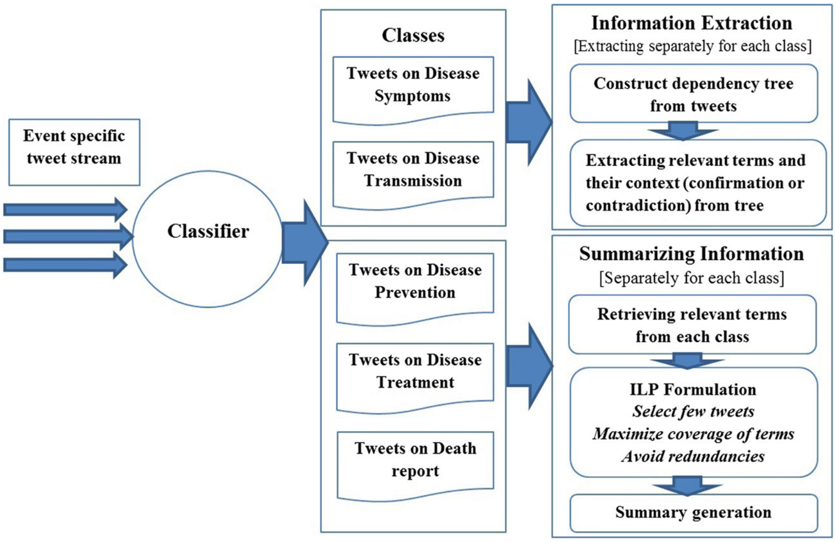

Event specific tweet stream
Classifier
Classes
Tweets on Disease Symptoms
Tweets on Disease Transmission
Tweets on Disease Prevention
Tweets on Disease Treatment
Tweets on Death report
Information Extraction
[Extracting separately for each class]
Construct dependency tree from tweets
Extracting relevant terms and their context (confirmation or contradiction) from tree
Summarizing Information
[Separately for each class]
Retrieving relevant terms from each class
ILP Formulation
Select few tweets
Maximize coverage of terms
Avoid redundancies
Summary generation

Fig. 1 Our proposed framework for classification-summarization of tweets posted during epidemic

2014; Centers for Disease Control and Prevention 2014) show that the proposed low-level lexical classifier outperforms vocabulary based approach (Imran et al. 2014) in cross-domain scenario (Section 4). Next, we show the utility of disease specific keyterms in capturing information from different disease related classes and summarizing those information (Section 5). We evaluate the performance of our proposed summarization scheme in Section 6. Our proposed ILP based summarization framework (MEDSUM) performs better compared to real time disaster summarization approach (COWTS) proposed by Rudra et al. (2015). Section 7 shows how extracted information satisfies needs of various stakeholders. Finally, we conclude our paper in Section 8.

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 (Kinnane and Milne 2010). Fox (2011) reported that 80% of internet users use online resources for information about health topics like specific disease or treatment. Further, it was shown that 34% of health searchers use social media resources to find health related topics (Elkin 2008). The popularity of social media in medical and health domain has gained attention from researchers for studying various topics on healthcare. This section provides a brief overview of various researches conducted for utilizing medical social media data in order to extract meaningful information and shows how they are different from traditional systems used for clinical notes.

2.1 Mining Information from Clinical Notes

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 (Friedman et al. 1999, 2004; Heinze et al. 2001; Hripcsak et al. 2002). The 2010 i2b2/VA challenge (Uzuner et al. 2011) presented the task of extracting medical concepts, tests and treatments from a given dataset. Most of the techniques follow Conditional Random Field (CRF) or rule based classifiers. Roberts et al. (Roberts and Harabagiu 2011) built a flexible framework for identifying medical concepts in clinical text, and classifying assertions, which indicate the existence, absence, or uncertainty of a medical problem. The framework was evaluated on the 2010 i2b2/VA challenge data. Recently Goodwin and Harabagiu (2016) utilized this framework for building a clinical question-answering system. They used a probabilistic knowledge graph, generated from electronic medical records (EMRs), in order to carry out answer inference.

2.2 Mining Health Information from Social Media Data

Scanfeld et al. (2010) used Q-Methodology to determine the main categories of content contained in Twitter users' status updates mentioning antibiotics. Lu et al. (2013) built a framework based on clustering analysis technique to explore interesting health-related topics in online health community. It utilized sentiment based features extracted from SentiWordNet (Esuli and Sebastiani 2007) and domain specific features from MetaMap. Denecke and Nejdl (2009) performed a comprehensive content analysis of different health related Web resources. It also classified medical weblogs according to their information type using features extracted from MetaMap. A framework based on Latent Dirichlet Allocation (LDA) to analyze discussion threads in a health community was proposed by Yang et al. (2016). They first extracted medical concepts, used a modified LDA to cluster documents and finally performed sentiment analysis for each conditional topic. Recently large scale researches have been done in exploring how microblogs can be used to extract symptoms related to disease (Paul and Dredze 2011), mental health (Homan et al. 2014) and so on.

Most of the methods proposed for extracting information from clinical text utilize earlier proposed systems (e.g., MetaMap (Aronson 2001), cTakes (Savova et al. 2010)) for mapping clinical documents to concepts of medical terminologies and ontologies (eg. UMLS (Bodenreider 2004), SNOMED CT (Stearns et al. 2001)). For a given text, these systems provide extracted terms concepts of clinical terminologies that can be used to describe the content of a document in a standardized way. However, tools like MetaMap were designed specifically to process clinical documents and are thus, specialized to their linguistic characteristics (Denecke 2014). The user-generated medical text from social media differs significantly from professionally written clinical notes. Recent studies have shown that directly applying Metamap on social media data leads to low quality word labels (Tu et al. 2016). There have also been works which propose methods for identifying the kind of failures MetaMap experiences when applied on social media data. Recently Park et al. (2014) characterized failures of MetaMap into boundary failures, missed term failures and word sense ambiguity failures.

Researchers also put lot of effort in designing text classification techniques (Imran et al. 2014; Rudra et al. 2015) suitable for microblogs. In our recent work, we propose a low-level lexical feature based classifier to classify tweets posted during epidemics (Rudra et al. 2017).

To our knowledge, all the existing methods try to extract knowledge from past medical records to infer solutions, diagnoses or treatment. However, these techniques will not work for sudden outbreaks for which past medical records are not available. There does not exist any real time classification-summarization framework to extract, classify, and summarize information from microblogs in real time. In this work, we take first step to this problem and propose a real time classification-summarization framework which can be applied to future epidemics.

3 Dataset and Classification of Messages

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

3.1 Epidemics

We collect the crisis-related messages using AIDR platform (Imran et al. 2014) 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 is 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, $\cdots$)

First, we remove non-English tweets using the language information provided by Twitter. After this step, we got around 200K tweets for MERS which were collected over a period of two and half months. However, tweets for Ebola were collected over a period of six months and we observe that most of the tweets (around 80%) posted after first two months are just exact or near duplicates of tweets posted during the first two months. Hence, for consistency, we select the first 200,000 tweets in chronological order for both the datasets. We make the tweet-ids publicly available to the research community at http://cse.iitkgp.ac.in/~krudra/epidemic.html.

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 employ human volunteers to observe different categories of disease tweets and to annotate them (details in Section 4). The disease categories identified by our volunteers (which agrees with prior works (Goodwin and Harabagiu 2016; Imran et al. 2016)) are as follows. Some example tweets of each category are shown in Table 1.

Disease-Related Tweets Tweets which contain disease related information are primarily of the following five types: (i) *Symptom* – reports of symptoms such as fever, cough, diarrhea, and shortness of breath or questions related to these symptoms. (ii) *Prevention* – questions or suggestions related to the prevention of disease or mention of a new prevention strategy. (iii) *Disease transmission* – reports

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)		
	Ebola	Early #ebola symptoms include fever headache body aches cough stomach pain vomiting and diarrhea
Symptom	MERS	Middle east respiratory syndrome symptoms include cough fever can lead to pneumonia & kidney failure
	Ebola	Ebola is a deadly disease prevent it today drink / bath with salty warm water
Prevention	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
	Ebola	Dozens flock to new liberia ebola treatment center new liberia ebola treatment center receives more than 100
Treatment	MERS	cn-old drugs tested to fight new disease mers
Death report	Ebola	The largest #ebola outbreak on record has killed 4000+
	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

of disease transmission or questions related to disease transmission. (iv) *Treatment* – questions or suggestions regarding the treatments of the disease. (v) *Death report* – reports of affected people due to the disease.

Non-disease Tweets Non-disease tweets do not contribute to disease awareness and mostly contain sentiment/opinion of people.

In this work, we try to extract information for both primary and secondary health care services. Symptom, prevention, and transmission classes are relevant to primary health care (vulnerable population) and information about treatment is necessary for secondary health care service (affected population). Finally, reports of dead and affected people are important for government and monitoring agencies.

The next two sections discuss our proposed methodology comprising of first categorizing disease-related information (Section 4), and then summarizing information in each category (Section 5).

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 consider 2000 randomly selected tweets (after removing duplicates and retweets) related to both the events. Three human volunteers independently observe the tweets, deciding whether they contribute to information about epidemic.[1] We obtain unanimous agreement (i.e., all three volunteers assign same label to a tweet) for 87% of the tweets. For rest of the tweets, we follow majority verdict. Non-disease category contains greater number of tweets as compared to tweets present in individual disease related 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 finally created.

4.2 Classification Features

We aim to build a classifier which can be trained over tweets posted during past disease outbreaks and can directly be used over tweets posted for future epidemics. Earlier Rudra

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

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

et al. (2015) 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 consider large medical knowledgebase, Unified Medical Language System (UMLS) (Bodenreider 2004). It comprises over 3 million concepts (virus, flu etc.), each of which is assigned to more than one of the 134 semantic types. Next, MetaMap (Aronson 2001) is used for mapping texts to UMLS concepts. For example, if MetaMap is applied over the tweet ‘Cover your mouth and wear gloves there is a mers warning’ then we get following set of word-concept type pairs—1. cover-**Medical device**, 2. mouth-**Body space**, 3. mers-**Disease or syndrome**, 4. gloves-**Manufactured object**, and 5. warning-**Regulatory activity**. As mentioned in Section 2, MetaMap does not perform well in case of short, informal texts. Hence, raw tweets have to pass through some preprocessing phases so that MetaMap can be applied over processed set of tweets. The preprocessing steps are described below.

1. We remove unnecessary words (URLs, mentions, hashtag signs, emoticons, punctuation, and other Twitter specific tags) from the tweets. We use a Twitter-specific part-of-speech (POS) tagger (Gimpel et al. 2011) to identify POS tags for each word in the tweet. Along with normal POS tags (nouns, verbs, etc.), this tagger also labels Twitter-specific elements such as emoticons, retweets, URLs, and so on.
2. We only consider words which are formal English words and present in an English dictionary (Aspell-python). We also remove out-of-vocabulary words commonly used in social media (Maity et al. 2016).
3. MetaMap is originally designed to work for formal medical texts. In case of general texts (tweets), we observe that many common words (‘i’, ‘not’, ‘from’) are mapped to certain medical concepts. For example, in the tweet `‘concern over ontario patient from nigeria with flu symptoms via’`, ‘from’ and ‘to’ are marked as *qualitative concept (qlco)*. Hence, we remove all the stopwords from tweets.

After preprocessing, tweets are passed as input to MetaMap which returns the set of tokens present in the tweet as concepts of UMLS Metathesaurus along with their corresponding semantic type. Finally, semantic types obtained from MetaMap are utilized for finding the relevant features. Table 3 lists the classification features (binary).

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 (‘sosy’); Physiologic Function (‘phsf’))
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

4.3 Performance

We compare the performance of our proposed set of lexical features with a standard bag-of-words (BOW) model similar to that in Imran et al. (2014) where unigrams are considered as features. We remove (URLs, mentions, hashtag signs, emoticons, punctuation, stopwords, and other Twitter-specific tags) from the tweets using Twitter pos tagger (Gimpel et al. 2011).

Model Selection For this experiment, we consider four state-of-the-art classification models from Scikit-learn package (Pedregosa et al. 2011)—(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 when our proposed set of features are used for training 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.

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. Values in the bracket represent standard deviations in case of in-domain accuracies

Train set	Test set			
	Ebola		MERS	
	BOW	PRO	BOW	PRO
Ebola	***84.78% (0.05)***	*84.02% (0.06)*	65.69%	**76.15%**
MERS	66.19%	**74.72%**	***88.26%(0.07)***	*81.05% (0.03)*

In-domain classification results are represented by italic entries. For each train-test pair, the accuracy of better performing system has been boldfaced

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 also measure precision, recall, F-score of classification for both sets of features. In order to take care of class imbalance, we consider weighted measure for precision, recall, and F-score. 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.

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. We observe that in most of the cases, tweets from ‘symptom’, ‘prevention’, and ‘transmission’ classes are incorrectly tagged as ‘non-disease’ due to absence of the features presented in Table 3. When we train our proposed model using Ebola dataset 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. A few pair of classes like ‘symptoms’ and ‘prevention’, ‘transmission’ and ‘prevention’, etc. are inter-related. In these pairs, people often use information from one class in order to derive information for the other class. Thus, we find simultaneous use of multiple classes in the same tweet. 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

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)*

In-domain classification results are represented by italic entries. For each train-test pair, the accuracy of better performing system has been boldfaced

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

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.

5 Summarization

Given the automatically classified tweets into different disease classes (described in previous section), in this section we aim to provide a cohesive summary of each class. The type of information and its representation to end-users that should be extracted from each category varies.[2] For instance, in the case of the 'symptom' category, two lists of symptoms are required (i) positive symptoms list (i.e. actual symptoms of a disease) (ii) negative symptoms list (i.e. symptoms which are not yet confirmed as actual symptoms of a disease). However, in the case of the 'prevention' category, instead of generating lists, we aim to summarize prevention strategies. Next, we describe different summarization techniques followed for different categories.

5.1 Summarizing Symptoms

To automatically extract positive and negative symptoms from the tweets classified into the symptom category, we first generate a symptoms dictionary. For this purpose, we extract symptoms listed on various credible online sources like Wikipedia,[3] MedicineNet,[4] Healthline[5] etc. Our dictionary contains around 770 symptoms.

Symptom Identification Now, given a tweet t, we check if it contains a symptom from the symptom dictionary. If a symptom s is found in t, then there can be two possibilities:

1. *Positive symptom*: The user who posted tweet t might be reporting that symptom s would be observed in a user if she is affected by the ongoing epidemic. Eg. 'symptoms of MERS include *fever* and *shortness of breath*.'
2. *Negative symptom*: The user who posted tweet t might be conveying that symptom s would not be observed if a user is affected by the ongoing epidemic. Eg. '#Ebola symptoms are different than upper respiratory tract pathogens, *no cough, nasal congestion* Dr. Wilson.'

We distinguish between the above two cases by using the terms having dependencies with the symptom term. We check if symptom s has a dependency (Kong et al. 2014) with any strongly negative term in the tweet t. Symptom s is a *negative symptom* of the disease if s has dependency with atleast one strongly negative term in t. If there is no such dependency with any negative term, then symptom s is a *positive symptom* of disease. We use Christopher Potts' sentiment tutorial[6] to identify strongly negative terms (e.g., never, no, wont) and Twitter dependency parser to identify the dependency relations present in a tweet. For example, in case of the tweet 'CDC announces second case of MERS virus.', the dependency tree returns following six relations — (CDC, announces), (case, announces), (second, case), (of, case), (MERS, virus), (virus, of). Table 7 shows examples of some positive and negative symptoms. After identifying both positive and negative symptoms, we try to rank them on the basis of their *corpus frequency* i.e., number of tweets in the corpus (symptom class) in which the symptom has been reported. However, the same symptom s

[2] https://www.cdc.gov/coronavirus/mers/about/symptoms.html,https://www.cdc.gov/coronavirus/mers/about/prevention.html

[3] https://en.wikipedia.org/wiki/List_of_medical_symptoms

[4] http://www.medicinenet.com/symptoms_and_signs/alpha_a.htm

[5] http://www.healthline.com/directory/symptoms

[6] http://sentiment.christopherpotts.net/lingstruc.html

Table 7 Sample tweets posted during outbreak containing symptoms in positive and negative context

Context	Tweet
Positive	#Ebola symptoms: fever, headache, muscle aches, weakness, no appetite, stomach pain, vomiting, diarrhea & bleeding
	RT @NTANewsNow: Ebola symptoms starts as malaria or cold then vomiting, weakness, Joint & Muscle Ache, Stomach pain and Lack of Appetite
	#Ebola symptoms are different than upper respiratory tract pathogens, no cough, nasal congestion Dr. Wilson
Negative	I've been informed that coughing is not a symptom of Ebola

might occur in multiple tweets. If a symptom s is found as a *positive symptom* in one tweet and also captured as a *negative symptom* in another tweet, then s is considered as ambiguous. Next, we describe the method to deal with such ambiguous symptoms.

Removing Conflicting Symptoms In this work, we are primarily interested in *positive symptoms* of a disease i.e. symptoms which represent that disease. As identified earlier, many ambiguous symptoms may occur in both positive and negative lists. However, the frequency of occurrence of a particular symptom s is not likely to be the same for both positive and negative classes. Hence, we compute the ratio of positive to negative corpus frequency of a particular ambiguous symptom s. If that ratio is ≤ 1, then we drop that symptom s from positive list.

5.2 Summarizing Disease Transmissions

During epidemics, vulnerable users look for information about possible disease transmission mediums so that precautionary steps can be taken. Common users and health organizations post tweets regarding possible transmission possibilities of a disease for public awareness. It is observed that information about transmission mediums is mostly centered around keywords like 'transmission', 'transmit' etc. In this work, we use following set of transmission related keywords—(i). transmission, (ii). transmit, (iii). transference, (iv). transferral, (v). dissemination, (vi). diffusion, (vii). emanation, (viii). channeling, (ix). spread, (x). transfer, (xi). relay.

To identify informative components centered around such keywords, we explore the dependency relation among the words in a tweet using a *dependency tree* (Kong et al. 2014). A dependency tree basically indicates the relation among different words present in a tweet. For example, dependency tree for the tweet 'Ebola virus could be transmitted via infectious aerosol' contains the following two dependency relations centered around keyword 'transmit'– (via, transmit), (aerosol, transmit). In general, the POS tag of every transmission medium will be 'Noun' (eg. 'aerosol' in the previous example). Hence, we detect all nouns connected to keywords in the dependency tree within a 2-hop distance.

It is observed that in some cases people post information about mediums not responsible for disease transmission. Table 8 shows example tweets providing information about transmission mediums in both positive and negative direction. To capture the actual intent of a message, we detect whether any negated context is associated with the keywords or not (same as proposed in symptom detection in Section 5.1). Finally, we rank the transmission mediums based on their *corpus frequency*, i.e., number of tweets in the transmission class in which they occur and remove ambiguous mediums (present in both positive and negative list) based on the ratio of their frequency of occurrence in positive and negative context (Section 5.1).

Table 8 Sample tweets posted during outbreak containing information about transmission mediums in positive and negative context

Context	Tweet
Positive	@USER @USER @USER I've also read that Ebola can spread thru airborne transmission [url]
	#Ebola virus could be transmitted via infectious aerosol particles
	Idiots & liars! @USER WH briefing: "Ebola is not like the flu. #Ebola is **not transmitted** through the air." [url]
Negative	RT @USER: CDc: You must have personal contact to contract #Ebola. It is **not transmitted** by airborn route

5.3 Summarizing Prevention Information

Users vulnerable to a disease are primarily looking for preventive measures. To provide a summary of those preventive measures, we take tweets categorized as 'preventive' by our classifier and some specific types of **preventive terms** which provide important information about preventive measures in epidemic scenarios—(i) Therapeutic or preventive procedure, (ii) symptom words, (iii). anatomy words (terms related to terms related to body system, substance, junction, body part etc). We extract these preventive terms from tweets using Metamap and UMLs knowledge bases.

Considering that the important preventive information in an epidemic is often centered around **preventive terms**, we can achieve a good coverage of preventive information in a summary by optimizing the coverage of such important preventive terms. In order to capture preventive terms, we extract prevention ('Therapeutic or preventive procedure'), anatomy('Body location or region', 'Body substance', 'Body part, organ, or organ component'), daily activity related concepts and terms using Metamap and UMLS. The importance of a preventive term is computed based on its frequency of occurrence in the corpus i.e., number of times a term t is present in the corpus of preventive tweets.

To generate a summary from the tweets in this category, we use an Integer Linear Programming (ILP)-based technique (Rudra et al. 2015) to optimize the coverage of the preventive terms. Table 9 states the notations used.

The summarization is achieved by optimizing the following ILP objective function:

$$max(\lambda_1.\sum_{i=1}^{n} x_i + \lambda_2.\sum_{j=1}^{m} Score(j).y_j) \qquad (1)$$

Table 9 Notations used in the summarization technique

Notation	Meaning
L	Desired summary length (number of words)
n	Number of tweets considered for summarization (in the time window specified by user)
m	Number of distinct content words included in the n tweets
i	Index for tweets
j	Index for preventive terms
x_i	Indicator variable for tweet i (1 if tweet i should be included in summary, 0 otherwise)
y_j	Indicator variable for preventive term j
$Length(i)$	Number of words present in tweet i
Score(j)	cf score of preventive term j
T_j	Set of tweets where content word j is present
P_i	Set of preventive terms present in tweet i

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \qquad (2)$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m] \qquad (3)$$

$$\sum_{j \in P_i} y_j \geq |P_i| \times x_i, i = [1 \cdots n] \qquad (4)$$

where the symbols are as explained in Table 9. The objective function considers both the number of tweets included in the summary (through the x_i variables) as well as the number of important preventive-terms (through the y_j variables) included. The constraint in Eq. 2 ensures that the total number of words contained in the tweets that get included in the summary are at most of the desired length L (user-specified) while the constraint in Eq. 3 ensures that if the preventive term j is selected to be included in the summary, i.e., if $y_j = 1$, then at least one tweet in which this preventive term is present is selected. Similarly, the constraint in Eq. 4 ensures that if a particular tweet is selected to be included in the summary, then the preventive terms in that tweet are also selected. In this summarization process, our objective is to capture more number of preventive terms rather than the number of tweets. Hence, λ_1 and λ_2 are set to 0 and 1 respectively.

We use GUROBI Optimizer (Gurobi 2015) to solve the ILP. After solving this ILP, the set of tweets i such that $x_i = 1$, represent the summary.

5.4 Summarizing Death Reports

During such epidemic apart from health related issues some socio-political matters also arise because travelers from foreign nations also get affected due to the ongoing epidemic and sometimes local government has to arrange necessary equipment for their treatment as well as send them back to their countries. Local residents suffering from the epidemic also need support from government and health agencies. Under such constraints government generally keeps track of number of people dead or under treatment. In this part, we try to extract this kind of information snippet from large set of tweets which may help government to get a quick snapshot of the situation.

Primarily, we observe that such information is centered around keywords like 'died', 'killed', 'dead', 'death', 'expire', 'demise' etc. Table 10 shows some examples of the tweets present in the 'death reports' class. While prior work (Rudra et al. 2015) considered all nouns and verbs as content words, in reality, all such keywords present in a tweet are *not* linked to health related events. Hence, in the present work, we identify the keywords for 'death reports' class from manually annotated corpus. As illustrated in Table 10,

Table 10 Sample tweets posted during outbreak containing information about killed or died people

As of Oct. 15th 2014 CDC numbers for #Ebola are 8997 total cases, 5006 laboratory-confirmed cases, and 4493 deaths in total
RT @USER: New WHO numbers on #Ebola outbreak in 3 West African countries: 1440 ill including 826 deaths. (As of 7/30)
#Ebola has infected almost 10,000 people this year, mostly in Sierra Leone, Guinea and Liberia, killing about 4900
RT @USER: #Ebola: As of 4 Aug 2014, countries have reported 1711 cases (1070 conf, 436 probable, 205 susp), incl 932 deaths

tweets contain location-wise information about dead people. Hence, it is necessary to capture location information in final summary. For summarization of death reports, we follow same ILP framework proposed in Section 5.3 but instead of optimizing **preventive terms**, here we optimize the coverage of **death related terms**. We consider numerals (e.g., number of casualties), keywords related to death reports, and location information as **death related terms**. We use Twitter-specific part-of-speech (POS) tagger (Gimpel et al. 2011) to identify POS tags for each word in the tweet. From these POS tags we select numerals for the summarization. We collect keywords related to death reports from manually annotated tweets. To identify location information, we use various online sources.[7] Finally, ILP method maximizes the coverage of death related terms.

5.5 Summarizing Treatment Information

Users who already get affected by the disease look for information about necessary medicines, treatment centers etc. For summarizing this information, we focus on tweets categorized as ‘treatment’ by our classifier and some specific types of **treatment terms** which provide important information about recovery procedure in epidemic scenario (i). clinical drug, (ii). pharmacologic substance (obtained from Metamap and UMLs). Table 11 provides examples of tweets containing treatment or recovery information.

Considering that the important recovery information in an epidemic is often centered around treatment terms, a good coverage of recovery information can be achieved by optimizing the coverage of important **treatment terms**. The importance of a treatment-term is computed based on its frequency of occurrence in the corpus i.e. number of times a treatment-term *t* occurs in the corpus of treatment related tweets. For summarization of treatment information, we follow the same ILP framework proposed in Section 5.3 but instead of optimizing preventive terms, here we optimize the coverage of **treatment related terms**.

Table 11 Sample tweets posted during outbreak containing recovery information

Fujifilm Drug Eyed As Possible Treatment For Ebola Virus
@USER Guarded optimism - use of #HIV antiviral to treat #ebola.
FDA-approved genital warts drug could treat #MERS
RT @USER: DNA vaccine demonstrates potential to prevent and treat deadly MERS coronavirus: Inovio Pharmaceuticals

We term our proposed MEDical dictionary based tweet SUMmmarization approach as **MEDSUM**. In the next section, we evaluate the performance of our proposed summarization models.

6 Experimental Results

In this section, we evaluate the performance of our proposed summarization techniques for different information classes (symptom, transmission, prevention, death information, treatment).

6.1 Evaluation of Symptoms of a Disease

In Section 5.1, we propose an algorithm to identify the symptoms of a disease. We need gold standard list of symptoms to check the accuracy of our method. We extract the actual symptoms of a disease by using online sources (World Health Organization (WHO) 2014; Centers for Disease Control and Prevention 2014) and compare the output of our algorithm with the actual symptoms to compute precision score. The number of actual symptoms for Ebola and MERS is 20 and 5 respectively. Hence, our proposed method also extracts 20 and 5 symptoms for Ebola and MERS respectively.

In case of Ebola, we observe that thirteen out of twenty symptoms are present in the gold standard list of symptoms. Three of the remaining seven symptoms are synonyms of some original symptom (present in the list of thirteen symptoms). Similarly, in case of MERS, three out of five symptoms are present in the gold standard list. Among the remaining two symptoms, one is synonym of some original symptom. Finally, our proposed method is able to extract sixteen and four original symptoms for Ebola and MERS respectively. Table 12 shows the precision and recall of our proposed approach.

Table 12 Precision and recall of our symptom identification method

Disease	Precision	Recall
Ebola	0.80	0.65
MERS	0.80	0.60

[7] https://en.wikipedia.org/wiki/Lists_of_cities_in_Africa, https://en.wikipedia.org/wiki/Middle_East

We observe that missed out symptoms (seven for Ebola and two for MERS) are identified at later stages of the disease which are not available in the tweets. Hence, we are not able to capture all the relevant symptoms but symptoms extracted from the tweets are able to give users an initial indication of the disease.

6.2 Evaluation of the Transmission Medium

In Section 5.2, we showed that users post information about both kinds of mediums i.e., mediums responsible for transmission (positive mediums) and mediums not responsible for transmission of the disease (negative mediums). Here, we are interested in positive transmission mediums i.e. mediums responsible for disease propagation. We extract the actual transmission mediums for both the diseases from online sources (World Health Organization (WHO) 2014; Centers for Disease Control and Prevention 2014) and compare the output of our algorithm with the actual transmission mediums to compute the precision and recall. We have collected fourteen and twelve transmission media for Ebola and MERS respectively. Table 13 shows the precision and recall of the proposed algorithm for top 10 and 20 transmission mediums.

It is clear from Table 13 that recall value will increase with more number of transmission mediums but precision goes down. However, many transmission mediums are identified at later stages which are not present in tweets posted during these epidemics. Still, it can provide a general overview about possible transmission mediums to the vulnerable people.

6.3 Evaluation of Prevention, Treatment Mechanisms and Death Reports

In case of symptom and transmission, we extract a ranked list of words and phrases from the tweets of corresponding classes. On the other hand, we propose an ILP based summarization scheme for prevention, death report, and treatment category. This method selects a set of tweets as a representative summary of the corresponding class. To measure the quality of system generated summary, we have to prepare ground truth summaries and compare system summaries with those ground truth summaries.

Table 13 Precision and recall of our transmission mediums detection method

Disease	#Mediums	Precision	Recall
Ebola	10	0.70	0.53
	20	0.65	0.92
MERS	10	0.50	0.42
	20	0.40	0.67

6.3.1 Experimental Settings

In the next part, we explain the baseline, evaluation criteria and results for each of the three information classes (prevention, death report, and treatment).

Preparing Ground-Truth Summaries For both the dataset and each of the information classes, three human volunteers (same as those involved in the classification stage) individually prepare summaries of length 200 words from the tweets of the corresponding class. To prepare the final ground truth summary of a particular disease and particular class, we first choose those tweets which are included in the individual summaries of all the volunteers, followed by those which are included by the majority of the volunteers. Thus, we create single ground truth summary of 200 words for each information class, for each dataset.

Baseline We compare the performance of our proposed summarization technique with disaster specific real time summarization technique COWTS proposed by Rudra et al. (Rudra et al. 2015).

Evaluation Metric We use the standard ROUGE (Lin 2004) metric for evaluating the quality of the summaries generated. Due to the informal nature of tweets, we actually consider the *recall and F-score* of the ROUGE-1 variant. Formally, ROUGE-1 recall is unigram recall between a candidate / system summary and a reference summary, i.e., how many unigrams of reference summary are present in the candidate summary normalized by the count of unigrams present in the reference summary. Similarly, ROUGE-1 precision is unigram precision between a candidate summary and a reference summary, i.e., how many unigrams of reference summary are present in the candidate/system summary normalized by the count of unigrams present in the candidate summary. Finally the F-score is computed as harmonic mean of recall and precision.

Next, we show the performance of our proposed method for each of the information classes.

6.3.2 Performance Comparison

Disease Prevention Table 14 gives the ROUGE-1 F-scores and recall values for both the algorithms respectively. It is clear that MEDSUM performs better compared to COWTS because disaster specific content words are not able to capture preventive information during disease outbreak.

Death Report Table 15 gives the ROUGE-1 F-scores and recall values for both the algorithms respectively. It is clear that our proposed method performs better compared to COWTS because disease related keywords capture more specific death related information compared to disaster specific content words.

Table 14 Comparison of ROUGE-1 recall and F-scores (Twitter-specific tags, emoticons, hashtags, mentions, urls, removed and standard rouge stemming(-m) and stopwords(-s) option) for MEDSUM (the proposed methodology) and the baseline method COWTS for prevention class

Event	MEDSUM		COWTS	
	Recall	F-score	Recall	F-score
Ebola	**0.4771**	**0.5195**	0.4575	0.5109
MERS	**0.4898**	**0.5393**	0.4761	0.4811

For each evaluation metric, the result of better performing system has been boldfaced

Disease Treatment Table 16 gives the ROUGE-1 F-scores and recall values for both the algorithms respectively. It is clear that coverage of treatment related information like drugs, medicines helps in better summarization.

In general, we observe that health related informative words are helpful to achieve better information coverage compared to disaster specific words during epidemics.

Further, we perform statistical t-test over six $(3(\#classes) * 2(\#datasets))$ ROUGE-1 F-scores (significance level 0.10) to check the statistical significance of MEDSUM over COWTS. The improvement appears to be statistically significant (the p-value is .0552).

7 Discussion

As stated earlier, primary objective of this work is to automatically extract and summarize information from microblog communications during epidemics to assist different stakeholders. In Section 5, we have proposed different summarization techniques for different information classes like 'symptom', 'prevention', 'treatment' etc. Next, we discuss how this information helps in primary and secondary health care service.

Table 15 Comparison of ROUGE-1 recall and F-scores (Twitter-specific tags, emoticons, hashtags, mentions, urls, removed and standard rouge stemming(-m) and stopwords(-s) option) for MEDSUM (the proposed methodology) and the baseline method COWTS for death reports

Event	MEDSUM		COWTS	
	Recall	F-score	Recall	F-score
Ebola	**0.4961**	**0.4980**	0.4961	0.4942
MERS	**0.3862**	**0.3758**	0.3448	0.3322

For each evaluation metric, the result of better performing system has been boldfaced

Table 16 Comparison of ROUGE-1 recall and F-scores (Twitter-specific tags, emoticons, hashtags, mentions, urls, removed and standard rouge stemming(-m) and stopwords(-s) option) for MEDSUM (the proposed methodology) and the baseline method COWTS for treatment class

Event	MEDSUM		COWTS	
	Recall	F-score	Recall	F-score
Ebola	**0.4803**	**0.4621**	0.3858	0.3525
MERS	**0.6517**	**0.5983**	0.4642	0.4244

For each evaluation metric, the result of better performing system has been boldfaced

Vulnerable Population Summarizing information for 'symptom', 'prevention', and 'transmission' classes helps assist vulnerable end-users and primary health care service. These communities are vulnerable to the disease and precautionary steps are extremely helpful to restrict further spreading of the disease. For example, if people are aware of possible transmission mediums (human-to-human, animal-to-human, air, aerosol etc) of the disease then they can avoid those possibilities and take relevant preventive measures.

Affected Population Post-disease community is mostly looking for treatment related information like hospital, drugs, medicines etc. In Section 5.5, we particularly tried to maximize such information via ILP approach. This kind of information helps in secondary health care services where treatment of patients is going on.

Health Organizations Government, health related organizations (WHO, CDC) looking for information about dead or affected people. Based on this kind of information they can decide whether medical response teams, new treatment centers etc. are necessary in certain regions or not.

Effect of Misclassification on Summarization As reported in Section 4, our proposed classifier is able to achieve around 80% accuracy in in-domain scenario and 75% accuracy in cross-domain scenario (25% tweets are classified wrongly). After classification, our proposed summarization framework summarizes the tweets present in the different disease related classes like 'symptoms', 'prevention' etc. In this part, we analyze the effect of misclassification on the summarization output. In the summarization, ILP framework tries to maximize the relevant class specific terms which represent a particular class. For example, in prevention category, ILP framework tries to maximize *prevention related terms*. If a prevention related tweet is misclassified then important terms present in that class are also wasted because such terms are not relevant to other classes (symptom, treatment etc.). Here, we measure the fraction of terms lost due to misclassification. For this analysis,

Table 17 Fraction of class specific terms covered and missed in symptom, prevention, and treatment class for both Ebola and MERS

Event	Symptom		Prevention		Treatment	
	Covered	Missed	Covered	Missed	Covered	Missed
Ebola	82.35	17.65	71.43	28.57	65	35
MERS	94.44	5.56	91.67	8.33	78.57	21.43

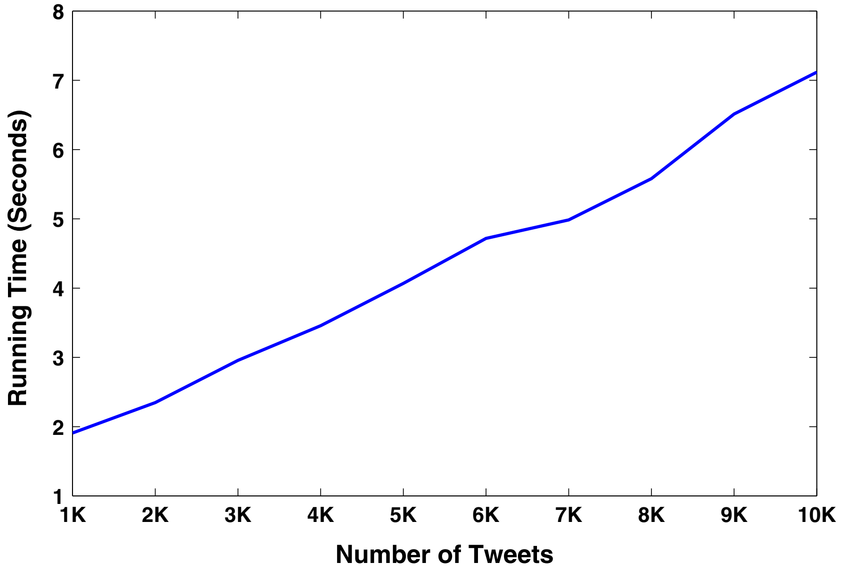
ground truth is required; hence, we measure these values over manually annotated ground truth data (Section 4). We consider two different cross-domain scenarios where the model is trained over Ebola and tested over MERS and vice versa. Table 17 shows the fraction of terms missed out for symptom, prevention, and treatment class for both Ebola and MERS. For MERS, we lose around 6–8% terms for symptom, prevention and 17% terms for treatment class. Similarly, for Ebola, around 17%, 28%, and 35% terms are lost due to misclassification for symptom, prevention, and treatment classes respectively. Overall, misclassification has an impact on overall summarization output. In future, we will incorporate other distinguishing low level features to reduce misclassification rate and improve the performance of classification-summarization framework.

Time Taken for Summarization During epidemic, it is necessary to summarize the information in real time because time is very critical in such scenarios. Hence, we analyze the execution time of various summarization approaches. For symptom and disease transmission, our proposed method takes around 172, and 257 seconds on average (over Ebola and MERS) respectively to generate summaries. For prevention, treatment, and death reports, proposed method takes around 7.39, 12.57, and 9.31 seconds respectively on average. Symptom and transmission mediums extraction take more time due to parsing overhead; still it is able to extract information in close to real time.

In this work, we observe that information is centered around some health related terms and we are trying to maximize the coverage of these terms in the final summary. We also measure the variation of running time with the number of tweets. We consider first 10,000 tweets from death report class of MERS (around 14,000 tweets) and measure the running time at ten equally spaced breakpoints (1000, 2000, $\cdots$, 10000). From Fig. 2, we can observe that running time increases more or less linearly with the number of tweets. However, number of terms which contain information during catastrophes or epidemics are less in number and also grow slowly compared to other real-life events like sports, politics, movies (Rudra et al. 2015). Hence, our proposed method is scalable and able to provide summaries in real time over large number of disease related tweets.

As most of the information during an epidemic is centered around some specific terms (prevention terms, drugs, treatment concepts), our proposed framework basically tries to maximize the coverage of these terms in the final summary. We believe that this framework may be extended to other crisis scenarios. However, health related terms (prevention terms, drugs, treatment concepts) will not work in those cases. We have to identify the terms which are capable of covering most of the important information during other kind of crisis scenarios.

Fig. 2 Variation of running time with number of tweets



8 Conclusion

Sudden disease outbreaks bring challenges for vulnerable and affected communities. They seek answers to their apprehensions; 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-summarization 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. Furthermore, various disease-category specific summarization approaches have been proposed. Often information posted on Twitter related to, for example, symptoms seems ambiguous for automatic information extractors. To deal with these issues, we generate separate lists representing positive and negative information. We make use of ILP techniques to generate 200-words summaries for some categories. Extensive experimentation conducted on real-world Twitter datasets from Ebola and MERS outbreaks show the effectiveness of the proposed approach. In future, we aim to deploy the system so that it can be practically used for any future epidemic.

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

Compliance with Ethical Standards

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

References

Aronson, A.R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA symposium* (p. 17). American Medical Informatics Association.

Aspell-python (2011). Python wrapper for aspell (C extension and python version). https://github.com/WojciechMula/aspell-python.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270.

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

De Choudhury, M. (2015). Anorexia on tumblr: a characterization study. In *Proceedings of the 5th international conference on digital health 2015* (pp. 43–50). ACM.

de Quincey, E., Kyriacou, T., Pantin, T. (2016). # Hayfever; a longitudinal study into hay fever related tweets in the UK. In *Proceedings of the 6th international conference on digital health conference* (pp. 85–89). ACM.

Denecke, K. (2014). Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study. In *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing*.

Denecke, K., & Nejdl, W. (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences*, *179*(12), 1870–1880.

Elkin, N. (2008). How America searches: health and wellness. Opinion Research Corporation: iCrossing pp. 1–17.

Esuli, A., & Sebastiani, F. (2007). SENTIWORDNET: a high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02 Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche Pisa IT.

Fox, S. (2011). *The social life of health information* Vol. 2011. Washington, DC: Pew Internet & American Life Project.

Friedman, C., Hripcsak, G., Shagina, L., Liu, H. (1999). Representing information in patient reports using natural language processing and the extensible markup language. *Journal of the American Medical Informatics Association*, *6*(1), 76–87.

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

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A. (2011). Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers* (Vol. 2, pp. 42–47). Association for Computational Linguistics.

Goodwin, T.R., & Harabagiu, S.M. (2016). Medical question answering for clinical decision support. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 297–306). ACM.

Gurobi (2015). The overall fastest and best supported solver available. http://www.gurobi.com/.

Heinze, D.T., Morsch, M.L., Holbrook, J. (2001). Mining free-text medical records. In *Proceedings of the AMIA symposium* (p. 254). American Medical Informatics Association.

Homan, C.M., Lu, N., Tu, X., Lytle, M.C., Silenzio, V. (2014). Social structure and depression in trevorspace. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 615–625). ACM.

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

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *Proceedings of the WWW companion* (pp. 159–162).

Imran, M., Mitra, P., Castillo, C. (2016). Twitter as a lifeline: human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European language resources association (ELRA), Paris, France*.

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

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., Smith, N.A. (2014). A dependency parser for tweets. In *Proceedings of the EMNLP*.

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (with ACL)*.

Lu, Y., Zhang, P., Deng, S. (2013). Exploring health-related topics in online health community using cluster analysis. In *46th Hawaii international conference on system sciences (HICSS), 2013* (pp. 802–811). IEEE.

Maity, S., Chaudhary, A., Kumar, S., Mukherjee, A., Sarda, C., Patil, A., Mondal, A. (2016). Wassup? lol: characterizing out-of-vocabulary words in twitter. In *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion, CSCW '16 companion* (pp. 341–344). New York: ACM.

Park, A., Hartzler, A.L., Huh, J., McDonald, D.W., Pratt, W. (2014). Automatically detecting failures in natural language processing tools for online community text. *Journal of Medical Internet Research*, *17*(8), e212–e212.

Paul, M.J., & Dredze, M. (2011). You are what you tweet: analyzing twitter for public health. *Icwsm*, *20*, 265–272.

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

Roberts, K., & Harabagiu, S.M. (2011). A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, *18*(5), 568–573.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., Ghosh, S. (2015). Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the CIKM*.

Rudra, K., Sharma, A., Ganguly, N., Imran, M. (2017). Classifying information from microblogs during epidemics. In *Proceedings of the 2017 international conference on digital health* (pp. 104–108). ACM.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507–513.

Scanfeld, D., Scanfeld, V., Larson, E.L. (2010). Dissemination of health information through social networks: twitter and antibiotics. *American Journal of Infection Control*, *38*(3), 182–188.

Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y. (2001). Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA symposium* (p. 662). American Medical Informatics Association.

Tu, H., Ma, Z., Sun, A., Wang, X. (2016). When metamap meets social media in healthcare: are the word labels correct? In *Information retrieval technology* (pp. 356–362). Springer.

Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L. (2011). 2010 I2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556.

World Health Organization (WHO) (2014). http://www.who.int/mediacentre/.

Yang, F.C., Lee, A.J., Kuo, S.C. (2016). Mining health social media with sentiment analysis. *Journal of medical systems*, *40*(11), 236.

Yom-Tov, E. (2015). Ebola data from the internet: an opportunity for syndromic surveillance or a news event? In *Proceedings of the 5th international conference on digital health 2015* (pp. 115–119). ACM.

Koustav Rudra received the B.E. degree in computer science from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, and the M.Tech degree from IIT Kharagpur, India. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India. His current research interests include social networks, information retrieval, and data mining.

Ashish Sharma is currently pursuing the Dual degree with the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India. His current research interests include social networks and information retrieval.

Niloy Ganguly received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, and the Ph.D. degree from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India. He was a Post-Doctoral Fellow with Technical University, Dresden, Germany. He is currently a Professor with the Department of Computer Science and Engineering, IIT Kharagpur, where he leads the Complex Networks Research Group. His current research interests include complex networks, social networks, and mobile systems.

Muhammad Imran is a Scientist at the Qatar Computing Research Institute (QCRI) where he leads the Crisis Computing team. His interdisciplinary research focuses on natural language processing, text mining, human-computer interaction, applied machine learning, and stream processing areas. Dr. Imran has published over 50 research papers in top-tier international conferences and journals. Two of his papers have received the Best Paper Award. He has been serving as a Co-Chair of the Social Media Studies track of the ISCRAM international conference since 2014 and has served as Program Committee (PC) for many major conferences and workshops including SIGIR, ICWSM, ACM DH, ICWE, SWDM. Dr. Imran has worked as a Post-Doctoral researcher at QCRI (2013-2015). He received his Ph.D. in Computer Science from the University of Trento, Italy (2013), where he also used to co-teach various computer science courses (2009-2012).