# A Robust and Scalable Framework for Detecting Self-reported Illness from Twitter

Muhammad Asif Hossain Khan
Graduate School of
Information Science and Technology
The University of Tokyo
Tokyo 153—8505, Japan
Email: asif@mcl.iis.u-tokyo.ac.jp

Masayuki Iwai
Institute of Industrial Science
The University of Tokyo
Tokyo 153—8505, Japan
Email: masa@iis.u-tokyo.ac.jp

Kaoru Sezaki
Center for Spatial Information Science
The University of Tokyo
Tokyo 153—8505, Japan
Email: sezaki@iis.u-tokyo.ac.jp

***Abstract*—Early detection of onset and outbreak of infectious diseases has paramount importance in containing such diseases before they turn into epidemics. The incredible growth in popularity and spatial resolution of coverage have made micro-blogging sites like Twitter a promising source of information for assessing the evolution of intensity of such diseases within a locality. However, identifying tweets with self-reported illness from other 'disease related' tweets is important for avoiding false alarms. In this research, our endeavor is to segregate the tweets all of which fall under the general category of 'disease related'. By using relatively very small training set and modifying the conventional $n$-gram feature selection method, we could isolate tweets reporting individual's illness with around 88.7% precision.**

***Index Terms*—Infodemiology, Trend analysis, Epidemic intelligence, Short text classification, Collective intelligence.**

## I. Introduction

Early detection of onset and outbreak of infectious diseases followed by timely intervention by authorities can greatly reduce the damage caused by these diseases. Most developed countries have their own disease surveillance systems, which unfortunately lag one week or more to present the data to the policy makers due to the massive data collection and compilation involved in the process. Hence, starting from Jhon Snows cholera study [1], much research have been focused on detecting outbreak of infectious diseases through alternative sources — in some cases by using differential equations and regression methods on simulated populations and hypothetical scenarios (*computational epidemiology*) [2], [3] and more recently by analyzing search engine queries [4], [5], and online social networks or micro-blog status updates (*infodemiology*) [6], [7], [8], [9], [10].

Recent research shows that Twitter posts can be used for capturing the overall trend of a particular disease outbreak [7], [8], [10]. The general assumption made in such research is that at the onset of an epidemic in an area, public concern will take a hike, which can be captured and quantified through the 'disease relevant' tweets generated from that locality. A set of disease related keywords are chosen, assigned weights and are considered as distinguishable features for identifying such tweets. For example, tweets containing terms like 'flu' or 'influenza' have generally been regarded as flu-relevant. However, our observation suggests that a significant proportion of tweets that consist of such important features do not report individual illness within the locality. We have identified three sources responsible for this confounding.

- All mainstream news media periodically update their news headlines. So, whenever there is a disease related event, whether it is a research breakthrough, an outbreak or a celebrity getting infected, Twitter is deluged with 'disease relevant' tweets.
- It can be observed that the use of bots for auto generation of advertisement tweets is commonplace nowadays. Advertisement for drugs and vaccines are 'disease relevant', however, have little importance for outbreak detection.
- The last and most difficult to discern are the tweets generated by individuals, which contain some important disease related keywords, but actually report no illness. We have discussed about this category in detail in the 'problem statement' section.

Any endeavor for monitoring the disease activity within a locality using Twitter could avoid raising false alarms if the aforementioned tweets could be segregated from the broad collection of disease relevant tweets.

In this research work, we have tried to determine whether a disease relevant tweet is reporting the author's illness or not. We have presented a method for classifying such tweets using a very small manually labeled training set of disease relevant tweets. We have shown that conventional $n$-gram feature selection methods do not ensure the best classification of such tweets. Our main contributions are as follows:

- We have developed a method for classifying tweets, all of which contain a common set of disease related keywords, into relevant and irrelevant classes. As these tweets share a set of keywords, the inter-class distance is quite marginal. Our method could still achieve acceptable accuracy and outperform the conventional $n$-gram feature based bag-of-word classifiers.
- Our proposed method can achieve acceptable accuracy in segregating misleading tweets using a very small training set. The method would spare the researchers in the field of *infodemiology* (information epidemiology) from the cost

of preparing a large training set and at the same time achieve better classification accuracy.

The rest of the paper is organized as follows. The problems we have identified and addressed for tweet classification have been delineated in section II. In section III we have discussed some methods adopted by contemporary researchers for classifying tweets and other short texts. Our proposed solution to the identified problems have been presented in detail in the forth section. In the fifth section we have discussed about the experiment data, different models that we have compared with our proposed model and some other evaluation details. Finally, we concluded this paper with a brief discussion about our future plans in the sixth section.

## II. Problem Statement

In this research we have focused our attention on the accurate identification of self-reported ILI (*Influenza Like Illness*) related tweets. Twitter imposes 140 character limit restriction on the length of each post — forcing its users to use abbreviations, non-grammatical sentences and non-dictionary words, sometimes followed by a link to a more elaborate source of information. This makes information extraction and classification of tweets particularly challenging. The length restriction on tweets incurs some unique research challenges as we have pointed out below.

- **Sparseness :** Being short, they do not contain enough distinguishable features, which makes their classification difficult. Moreover, to compensate the brevity per document (tweet), it is necessary to incorporate thousands of tweets per class in the training data set. Manual annotation of such huge corpus is not a feasible option. Though some researchers have used *Amazon Mechanical Turk* for labeling a few thousand tweets for their research, it is expensive, subject to human error and cannot be adopted by many due to the incurred cost. A cursory labeling of a few hundred tweets for building a model to automatically get rid of the noisy tweets would help a lot of researchers and that is what we have tried to achieve through this research.
- **Ambiguity:** As authors of tweets cannot clearly express themselves due to the imposed length restriction, there is often insufficient evidence present in the text of the tweet to distinguish relevant tweets from irrelevant. Consider the tweet "Get the flu *with the shot!! But I never* have the flu *... So, I dont want the shot ...*". Here the author is referring to a promotional slogan for some flu vaccine in his first sentence and his opinion about that in the following sentences. Most self-reported flu tweets contains the sentences like *"I have the flu ..."* or *"I got the flu ..."* Hence, due to the presence of the two phrases "*get the flu*" and '*have the flu*" in the first tweet, most bag-of-word based classifiers trained with conventional $n$-gram features would confuse it for a *self-reported flu* tweet. We have addressed this problem in our research.
- **Connotation:** It is very difficult to pick the author's tone from a 140 character string. Let us consider the tweets *"I wish I could say 'I will miss the class due my anxiety' rather than lying that 'I m having flu' ..."* or *"If I have one more flu, I can reach my goal weight ..."*. Some researchers like Krieck *et al.* [6] have ruled out any tweet having an interrogation sign or starting with a conditional word (e.g. *if*) to be irrelevant. However, we have encountered many relevant tweets having these constructs. For example, the tweets *"If this damn flu doesnt leave me by tomorrow, I am gonna kill myslf ..."* or *"How long do I have to suffer from this flu? ..."* are definitely positive examples of self-reported illness.

We have tried to address the problem of sparseness and ambiguity in this research work. To make the classification problem even more challenging, we have tried to classify tweets all of which contains either one or both of the most common ILI related keyword — 'flu' and 'influenza'. We have tried to classify them into three classes:

1) **Self:** The tweets in which authors express their flu infection either directly or indirectly. Example tweets from this class are *"In bed all day with the flu. Not fun ..."* or *"Drinking Thera Flu to try to Beat this ... wish someone were here taking care of me ..."*.
2) **News:** This is the class of tweets that reports some events regarding ILI. They may flood the Twitter stream when news of epidemic breaking out anywhere in the world is reported by mainstream media. However, most of such tweets are actually *Retweets*, which are tweets forwarded by a Twitter user, but not created by him/herself. Fortunately, Twitter provides mechanism to identify Retweets. Still we have encountered significant proportion of tweets reporting some news that are not retweets. Example of tweet belonging to this class is *"Chinese bus driver infected with H1N1 bird flu virus dies; Country's first reported human case in 18 months ..."*.
3) **False:** These are the tweets that contain the disease related keywords, but do not fall into any of the above categories. For example, in the tweet *"In bed ... istening to DJ Khaled's Thera Flu ..."* the user is referring to a popular music track, not any illness. Please notice the similarity between this tweet and the first example tweet of class 'self'. Isolating these tweets are particularly challenging and this is where the conventional $n$-gram feature based classifiers suffer most.

## III. Related Work

The problem of classifying tweets into relevant and irrelevant groups falls under the general problem of *short text classification*. There has been a number of different approaches adopted by different researchers in the filtering step. Lampos *et al.* [8] tried two different approaches for selecting a set of $n$-grams based on which they decided whether a tweet is flu related or not. In their first approach they hand-selected a set of 41 unigrams and bigrams and by using a least square linear regression algorithm they determined weights of different keywords. In our research we tried to avoid hand selection

of keywords and tried to learn best $n$-gram features from the tweets themselves. Considering the difference of predilection for words by Twitter users from different regions, it seems more reasonable to use different set of keywords for different regions. In their second approach, they collected 1560 disease and symptom related keywords from web articles related to influenza and from Wikipedia. Using a linear optimization model, they selected the final subset of 97 keywords that best correlates with the official HPA (*Health Protection Agency* of UK) flu rates. Again, selecting keywords based on some different domain like Wikipedia, which imposes no length limitations on articles, has possibility of facing the problem of *domain adaptation*.

Culotta [7] used a bag-of-word classifier to identify ILI related tweets. The training set was a set of 206 tweets containing one or more of the five keywords hand-selected by the author and manually classified as influenza related or not. Ritterman *et al.* [10] showed that Twitter data is a valuable information channel for predicting public opinion regarding the likelihood of a pandemic. They observed the frequency of a set of 1431 unigrams and 347 bigrams in the daily twitter corpus and used a support vector machine algorithm to carry out regression. The $n$-grams were selected based on their frequency of occurrence in the experiment data. Selecting bigram features based on frequency of occurrence do not guarantee avoiding the NLP problem called *occurrence-by-chance*, which we have addressed in our research and have discussed in length in the 'proposed model' section.

Some researchers have tried to use unsupervised learning methods for classifying short text. Pozdonoukhov *et al.* [12], who have tried to identify occurrence of social events using tweets generated from Ireland, have used a variant of the popular unsupervised topic model LDA [13], called streaming LDA for clustering tweets of similar topic. Then, they have analyzed key terms representing each identified topic to find the topic of their interest and have used tweets classified under that topic for making predictions. Phan *et al.* [14] have tried to classify short text like web snippets returned by search engines. To compensate for the sparseness of the snippet they appended to each snippet the topics hidden in the snippet. They have used LDA trained on Wikipedia data to identify the hidden topics. Topic identification does not fit our problem in hand as we are trying to filter tweets relevant to a single topic, which is a particular disease, in our case influenza.

There have been some efforts for classifying short text based on the presence of some fixed phrases. Bollen *et al.* [11] have tried to predict the closing values of Dow Jones Industrial Average through mood analysis of Twitter messages generated from New York area. In the tweet classification step they have filtered out tweets that contains explicit statement of the author's mood states, e.g. "i am feeling", "i dont feel", "makes me" etc. Instead of manually determining the phrases, in our model we are learning the collocation features from the training set. Moreover, we are using a variable collocation window, which helps us to capture more versatile collocation structures.

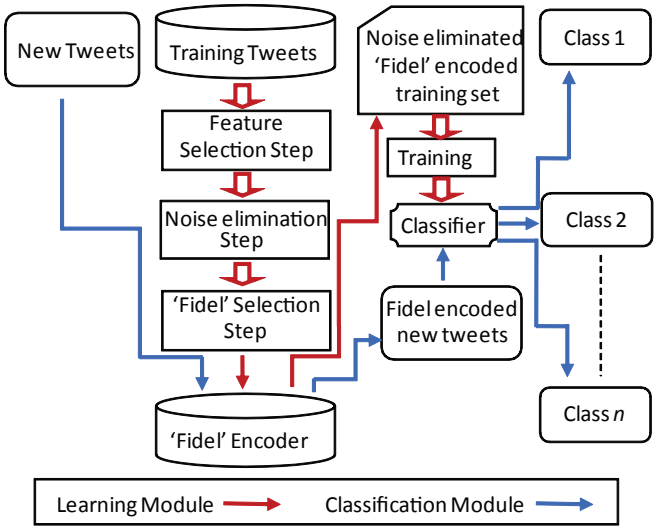


Fig. 1. General framework of our proposed method with learning and classification modules

## IV. PROPOSED MODEL

We have proposed a tweet classification model for identifying self-reported illness. Figure 1 shows the framework of our proposed model. The *Learning Module* (red lines) depicts our solutions to the following sub-problems:

- Identifying the most important collocation features based on likelihood ratio from training set
- Getting rid of noisy tweets from training set
- Building the classifier based on the selected features

The *Classification Module* (blue lines) is responsible for classification of further tweets. In the following subsections we have discussed the submodules in detail.

### A. Identifying the Most Significant Unigram and Bigram Features

*1) Capture Bigrams with Flexible Structure:* Many collocations consists of two words whose association follows a flexible relationship. The two words comprising the bigram do not appear in a fixed distance from each other. For example, in the excerpts *"I* feel sick *today . . . "*, *"I* feel *so* sick *today . . . "*, *"If I* feel *a bit more* sick *. . . "*, the words in the bigram *"feel sick"* appear in distance 1, 2 and 4 respectively. To capture such versatile structure, we have used a collocation window of 4 and have considered every word pair within the window as a potential collocation bigram. For example, the phrase "powerful personal computer" would produce three bigrams; 'powerful personal', 'personal computer' and 'powerful computer' when the collocation window is set to 2 or a higher number.

*2) Avoid Collocation by Chance:* The simplest way of finding significant bigram features is to select most frequently occurring bigrams. However, when the training corpus is small, two words might co-occur a lot just by chance. For example, if the training corpus were the set of those tweets generated from New York during the first week of April 2012

that contain the word 'flu', the bigram '*west flu*' will appear with significantly high frequency. The reason is that during that time a new music album was released by the popular singer 'Kanye West' and the title of one of the tracks was '*Thera flu*'. Considering '*west flu*' as a potentially important bigram feature for identifying 'flu related tweets' would be a mistake. To determine whether the bigram has some real structural importance, we have adopted the 'Likelihood Ratio' approach for hypothesis testing of independence, which takes into account the volume of data that has been considered for calculating the frequency of the bigram as well as the frequency of the individual words comprising the bigram. For sparse data this approach is more appropriate than the $\chi^2$ test [15].

Likelihood ratio is a number that tells how much more likely one hypothesis is over another. Our first hypothesis $H_1$ states that there is no association between the words beyond chance occurrence; i.e. in the bigram $w^1w^2$, the words $w^1$ and $w^2$ are generated completely independently of each other. The second hypothesis $H_2$ states that there is a structural dependence between $w^1$ and $w^2$. Formally,

**Hypothesis 1** $(H_1)$**:** $P(w^2|w^1) = p = P(w^2|\neg w^1)$

**Hypothesis 2** $(H_2)$**:** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

We have used the usual maximum likelihood estimates for calculating $p, p_1$ and $p_2$. Let, $c_1, c_2$ and $c_{12}$ be the number of occurrences of $w^1, w^2$ and $(w^1, w^2)$ in the text corpus respectively. The likelihood of getting the counts $w^1, w^2$ and $(w^1, w^2)$ in the current corpus is

$L(H_1) = bin(c_{12}; c_1, p)bin(c_2 - c_{12}; N - c_1, p)$ and

$L(H_2) = bin(c_{12}; c_1, p_1)bin(c_2 - c_{12}; N - c_1, p_2)$.

Here, $bin(x; n, p) = \binom{n}{x}p^x(1-p)^{n-x}$ represents binomial distribution. We then calculate the likelihood ratio $\lambda = \frac{L(H_1)}{L(H_2)}$ of the two hypotheses. The quantity $(-2\log\lambda)$ is asymptotically $\chi^2$ distribution. We reject the hypothesis of independence $(H_1)$ for a bigram with 95% confidence if $-2\log\lambda \geq 7.88$, which is the critical value for $\chi^2$ distribution with 1-degree of freedom and a confidence level of $\alpha = 0.005$.

*3) Determine Most Appropriate Class for Overlapping bigrams:* When the same bigram appears in the training set for more than one class and passes the test of 'collocation by chance', we tried to determine the class for which the bigram is more appropriate as a characteristic feature. For such bigrams, we checked the 'ratios of relative frequencies' between two or more classes to determine the most appropriate class for the the bigram. Let, $c_1$ and $c_2$ be the frequencies of a bigram $b$ in the training corpus of classes $X$ and $Y$ respectively. Let, $N_1$ and $N_2$ are total number of bigrams identified from classes $X$ and $Y$. Then the relative frequency ratio $r = \frac{c_1/N_1}{c_2/N_2}$. If $r \geq 1$, then $b$'s most appropriate class is $X$, otherwise it is $Y$.

*4) Unigram Feature Selection:* For unigram features, we have used the $\chi^2$ feature selection method.

### B. Tweaking the Noisy Tweets

To address the problem of ambiguity, as discussed in the *problem statement* section, we have developed Algorithm 1 to identify the noisy tweets. For each class $i$, we first retrieve the set of bigrams that are identifying features for this class (lines 2 — 5). These are the bigrams selected by the methods described in the previous sub-section. For every tweet in the training corpus of a class $i$, we determine the number of bigrams in the tweet that are identifying features for that class (lines 9 — 16). If less than 50% of the bigrams in a tweet are not identifying features for the class it belongs to, we discard that tweet from the training set (lines 17 — 19).

**Algorithm 1** Remove Noisy Tweets from Training Set

```
1.  allBigrams, trainingSet, fidels ← null
2.  for each class i do
3.      bigramsOf[i] ← identifying feature bigrams of class i
4.      allBigrams ← allBigrams ∪ bigramsOf[i]
5.  end for
6.  for each class i do
7.      for each tweet j ∈ i do
8.          global ← 0 local ← 0
9.          for each bigram b ∈ j do
10.             if b ∈ allBigrams then
11.                 global ← global + 1
12.             end if
13.             if b ∈ bigramsOf[i] then
14.                 local ← local + 1
15.             end if
16.         end for
17.         if local/global < 0.5 then
18.             trainingSet ← trainingSet ∪ {j}
19.         end if
20.     end for
21. end for
```

### C. Refining the Selected Bigrams

Among the bigrams, those which are still present in at least one of the tweets in the training set after the tweaking step, form the set of final bigram features. For the convenience of reference, we name this set '***fidels***'. Algorithm 2 identifies the *fidels*.

**Algorithm 2** Identify Fidels

```
1. for each tweet t ∈ trainingSet do
2.     for each bigram b ∈ t do
3.         if b ∈ allBigrams then
4.             fidels ← fidels ∪ {b}
5.         end if
6.     end for
7. end for
```

## V. Evaluation

### A. Experiment Data

We have used Twitter's 'search API' to develop a Java based crawler and have been crawling tweets from New York, London and Boston at an average rate of 200,000 tweets per day from each city. For this experiment we have used tweets generated in NY from December 06, 2011 to April 30, 2012. We have only considered those tweets which contains of either

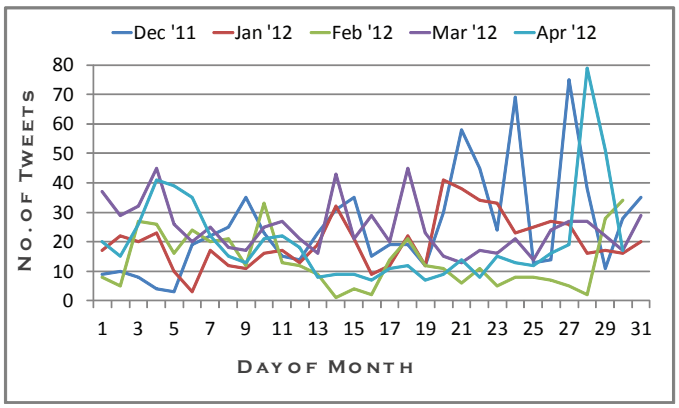


Fig. 2. Monthly distribution of tweets having the keywords 'flu' or 'influenza'

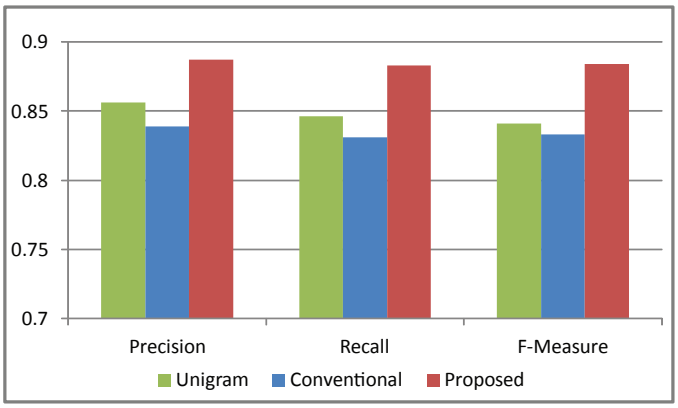


Fig. 3. Comparison among average performance among the three models

the keywords 'flu' or 'influenza'. A total of $3,955$ tweets had these keywords and we randomly selected $887$ tweets from this corpus. The monthly distribution of the tweets are shown in figure 2. We have then manually labeled them into three classes: self (329), news (268) and false (290). We applied the following pre-processing to all the training tweets:

- Removed all punctuation other than sentence's end markers.
- Removed mention of other Twitter users
- Removed all hashtags and urls.
- Converted everything to lowercase.
- Replaced repetition of the same character more than twice with a sequence of two characters; e.g. 'happppy' became 'happy'.

### B. Models Considered for Comparison

We have compared the performance of our proposed model with two other models. The difference among the models are in the set of features they consider during the learning phase.
**Unigram model:** Considers only unigram features.
**Conventional model:** Considers unigram and bigrams features selected though $\chi^2$ feature selection and likelihood ration respectively.
**Our proposed model:** Considers unigram features ($\chi^2$ method) and *fidels*.

### C. Evaluation Metrics

In the evaluation phase we aim to compare the performance of a text classifier when trained with features selected by our algorithm versus those selected by conventional NLP methods. We have used precision, recall and F-measure for comparing the performances of the different models. *Precision* is the fraction of retrieved tweets that are relevant and is defined as $P = TP/(TP+FP)$. *Recall* is the fraction of relevant tweets that are retrieved and is defined as $R = TP/(TP + FN)$. The terms $TP, TF, FP$ and $FN$ are defined in table 1. Both precision and recall are important measures in the the field of information epidemiology. High precision makes sure that irrelevant tweets will not convolute the conclusion reached at inference step, whereas, a good recall ensures most relevant tweets are being taken into consideration. *F-measure*, also known as $F_1 - score$, is the harmonic mean of precision and recall and is a convenient way for measuring the classification performance using a single numeric value. It is defined as, $F = 2 * \frac{precision*recall}{precision+recall}$.

TABLE I
TERM DEFINITION

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | True Positive ($TP$) | False Positive ($FP$) |
| Not Retrieved | False Negative ($FN$) | True Negative ($TN$) |

### D. Classifier

We have used a multinomial Naïve Bayes classifier for classifying the tweets into the three classes. Multinomial Naïve Bayes is a probabilistic learning method. The probability of a tweet $d$ being in the class $c$ is computed as:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(f_k|c)$$

where $P(f_k|c)$ is the conditional probability of feature $f_k$ occurring in a tweet of class $c$ and $n_d$ is the number of features encountered in tweet $d$. $P(c)$ is the prior probability of a tweet occurring in class $c$, which is obtained through maximum likelihood estimates. The best class for tweet $d$ is the *maximum a posteriori* (MAP) class $c_{map}$:

$$c_{map} = argmax_{c \in \mathbb{C}} \hat{P}(c|d) = argmax_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(f_k|c)$$

We have used *Laplace smoothing* for accommodating unseen feature space. 10-fold cross validation method has been adopted for assessing the classification performance.

### E. Result and Discussion

Figure 3 shows the overall performance comparison among the three models. The proposed model shows better precision and recall compared to both the conventional and unigram

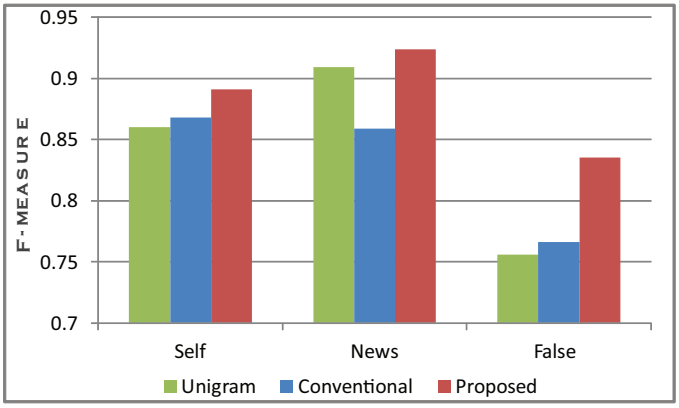


Fig. 4. Per class performance comparison among the three models

models. The performance improvement due to the 'tweaking of noisy tweets' and 'refined bigram selection' becomes evident when the precision, recall or F-measure of the conventional model is compared to that of the proposed model. This is because both these models are using both unigram and bigram features. The reason why the overall performance of unigram model outperforms the conventional model is describe in the next paragraph with the help of figure 4.

Figure 4 depicts the per-class classification performance represented in F-measure. As mentioned earlier, the main objective of this research work is to segregate self-reported illness from other 'disease related' tweets. The improved F-measure of the proposed model compared with the other two models for all three classes substantiates the effectiveness of our proposed algorithms. As it can be seen from Figure 4, the conventional model performs slightly better than the unigram model for classes 'Self' and 'False'. However, for the class 'News' unigram model outperforms the conventional model with significant margin due to which the overall performance of the conventional model is suffers (figure 3). Out of the 268 'News' tweets, unigram model correctly classified 244, whereas, the conventional model could correctly classify only 213. 45 of the 'News' tweets had been classified as 'False'. This is because, many of the 'News' tweets shared the bigram feature of class 'False'. In our proposed method, we could identify and remove these features to prevent them from convoluting the judgment of the classifier.

## VI. Conclusion and Future Work

Early identification of the onset of any infectious disease can prevent large scale outbreak of such diseases. Several previous research have shown that online social networks and micro-blogging sites can be useful source of information for picking up early hits about the onset of such diseases. However, we have to be cautious not to raise false alarms, which would incur unnecessary cost and create widespread panic in the society. Especially, when the source of information is Twitter, which is notoriously noisy, extra precautions are well justified. We have observed that a significant proportion of tweets which contain many of the disease related keywords actually do not report any sort of illness. Hence, taking these tweets into consideration for making predictions about the level of infection in a locality has every possibility of crying wolf. In this paper, we have proposed algorithms to get rid of such misleading tweets. Our proposed method could achieve 88.7% precision in classifying tweets with self-reported illness from other two disease related tweet classes that we have identified. Moreover, our method can work with relatively few manually labeled tweets. Thus, it could be easily scalable when the number of classes would increase.

We are working on some interesting extensions of this work. As mentioned in the *problem statement* section, in this research work we have dealt with two out of the three problems we have identified in classifying disease related tweets. We are now working on dealing with the problem of 'connotation'. We are integrating a sentiment analysis part in our framework to deal with the problem. We shall also try higher order $n$-grams. Moreover, we would like to assign different weights to the identified features to obtain even better classification accuracy. We hope that the inclusion of these extensions would make our model a practical and faster alternative to the conventional disease surveillance systems.

## References

[1] J. Snow, "On the mode of communion of cholera", *Jhon Churchill*, Ch 5, 1855.
[2] S. Eubank, H. Guclu, V. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang, "Modelling disease outbreaks in realistic urban social networks", *Nature*, vol. 429(6988) , pp. 180-184, 2004.
[3] B. Grenfell, O. Bjornstad and J. Kappey, "Travelling waves and spatial hierarchies in measles epidemics", *Nature*, vol. 414(6865), pp. 716-723, 2001.
[4] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Bramer M. Smolinksi and L. Brilliant, "Detecting influenza epidemics using search engine query data", *Nature*, vol. 457(7273), pp. 1012-1014, 2008.
[5] P.M. Polgreen, Y. Chen, D.M. Pennock and F.D. Nelson, "Using Internet Searches for Influenza Surveillance", *Clinical Infectious Diseases*, vol. 47(11), pp. 1443-1448, 2008.
[6] M. Krieck, J. Dreesman, L. Otrusina and K. Denecke, "A new age of public health: Identifying disease outbreaks by analyzing tweets", In *Proc. of HWS Workshop, ACM Web Science Conf.*, 2011.
[7] A.Culotta, "Towards detecting influenza epidemics by analyzing Twitter message", In *Proc. of the First Workshop on Social Media Analytics*, pp. 115-122, 2010.
[8] V. Lampos, T.D. Bie and N. Cristianini, "Flu detector – tracking epidemics on Twitter", *Machine Learning and Knowledge Discovery in Databases* , pp. 599-602, 2010.
[9] R. Chunara, J. Andrews and J. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak", *The American Journal of Tropical Medicine and Hygiene* , vol. 86(1), pp. 39-45, 2012.
[10] J. Ritterman, M. Osborne and E. Klein, "Using prediction markets and Twitter to predict a swine flu pandemic", *In Proc. of 1st Intl. Workshop on Mining Social Media*, 2009.
[11] J. Bollen, H. Mao, and X. J. Zeng, "Twitter mood predicts the stock market", *Journal of Computer Science*, vol. 2-1, pp. 1-8, Mar. 2011.
[12] A. Pozdonoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text", In *Proc. of ACM LBSN*, Chicago, IL, USA, Nov. 2011.
[13] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation", *J. Mach. Learn. Res.*, vol. 3, pp. 993-1022, 2003.
[14] X. H. Phan, L.M. Nguyen and S. Horiguchi, "Learning to classify short and sparse text and web with hidden topics from large-scale data collections", In. *Proc. of WWW*, pp. 91-100, Beijing, China, 2008.
[15] C.D. Manning and H. Schutze, "Foundations of Statistical Natural Language Processing", *The MIT Press*, Ch 5, 1999.