Journal of Biomedical Informatics 49 (2014) 255–268

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages

Suppawong Tuarob [a], Conrad S. Tucker [b,a,*], Marcel Salathe [c], Nilam Ram [d]

[a] Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[b] Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA
[d] Human Development and Family Studies, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 15 October 2013
Accepted 6 March 2014
Available online 16 March 2014

Keywords:
Social media
Machine learning
Classification

ABSTRACT

Objectives: The role of social media as a source of timely and massive information has become more apparent since the era of Web 2.0. Multiple studies illustrated the use of information in social media to discover biomedical and health-related knowledge. Most methods proposed in the literature employ traditional document classification techniques that represent a document as a bag of words. These techniques work well when documents are rich in text and conform to standard English; however, they are not optimal for social media data where sparsity and noise are norms. This paper aims to address the limitations posed by the traditional bag-of-word based methods and propose to use heterogeneous features in combination with ensemble machine learning techniques to discover health-related information, which could prove to be useful to multiple biomedical applications, especially those needing to discover health-related knowledge in large scale social media data. Furthermore, the proposed methodology could be generalized to discover different types of information in various kinds of textual data.
Methodology: Social media data is characterized by an abundance of short social-oriented messages that do not conform to standard languages, both grammatically and syntactically. The problem of discovering health-related knowledge in social media data streams is then transformed into a text classification problem, where a text is identified as positive if it is health-related and negative otherwise. We first identify the limitations of the traditional methods which train machines with *N*-gram word features, then propose to overcome such limitations by utilizing the collaboration of machine learning based classifiers, each of which is trained to learn a semantically different aspect of the data. The parameter analysis for tuning each classifier is also reported.
Data sets: Three data sets are used in this research. The first data set comprises of approximately 5000 hand-labeled tweets, and is used for cross validation of the classification models in the small scale experiment, and for training the classifiers in the real-world large scale experiment. The second data set is a random sample of real-world Twitter data in the US. The third data set is a random sample of real-world Facebook Timeline posts.
Evaluations: Two sets of evaluations are conducted to investigate the proposed model's ability to discover health-related information in the social media domain: small scale and large scale evaluations. The small scale evaluation employs 10-fold cross validation on the labeled data, and aims to tune parameters of the proposed models, and to compare with the stage-of-the-art method. The large scale evaluation tests the trained classification models on the native, real-world data sets, and is needed to verify the ability of the proposed model to handle the massive heterogeneity in real-world social media.
Findings: The small scale experiment reveals that the proposed method is able to mitigate the limitations in the well established techniques existing in the literature, resulting in performance improvement of 18.61% (*F*-measure). The large scale experiment further reveals that the baseline fails to perform well on larger data with higher degrees of heterogeneity, while the proposed method is able to yield reasonably good performance and outperform the baseline by 46.62% (*F*-Measure) on average.

* Corresponding author.
E-mail address: ctucker4@psu.edu (C.S. Tucker).

http://dx.doi.org/10.1016/j.jbi.2014.03.005
1532-0464/

1. Introduction

Social media such as Twitter and Facebook are increasingly being used as tools for real-time knowledge discovery relating to social events, emerging threats, epidemics, and even product trends [1,2]. For example, real time analysis of Twitter users' tweet content can be or is being used to detect earthquakes and provide warnings [3], to identify needs (e.g., medical emergencies, food and water shortages) during recovery from natural disasters such as the Haiti Earthquake [4], track emergence of specific syndromic characteristics of influenza-like illness [5], and collect epidemic-related tweets [6].

The role of social media in biomedical domain has become significant in recent years [7–13]. Researchers and physicians have utilized social media data to (1) communicate and share information between patients and health care decision makers, (2) develop large scale, dynamic disease surveillance systems and (3) mining biomedical and health-related information.

An immediate and direct use of social media in the biomedical domain is a means for patients and professionals to communicate and exchange information. Web 2.0 along with ubiquitous mobile computing devices allows individuals to dynamically and seamlessly interact with each other in real time, regardless of their locations. PatientsLikeMe[1] is a social network for patients that improves lives and a real-time research platform that advances medicine. On PatientsLikeMe's network, patients connect with others who have the same disease or condition, allowing them to track and share their own experiences. Eijk et al. illustrated the use of Online Health Communities (OHCs) for ParkinsonNet,[2] a social network for Parkinson disease patients whose participants (both patients and professionals) use various types of OHCs to deliver patient-centered care [14]. Merolli et al. explored different ways that chronic disease sufferers engage in social media in order to better tailor these online interventions to individually support patients in specific groups [9]. Additionally, Twitter, Facebook, and other social blogging services provide conduits for patients and medical practitioners to collaborate, exchange, and disseminate information through official broadcasting channels/webpages or discussion groups [12,10,13,11,15].

Most popular social media providers such as Twitter and Facebook allow their posts to be geo-located. These properties provide researchers in the healthcare community the ability to monitor the medical related emergences. Culotta proposed a methodology to study the predictability of Twitter data on future influenza rates [16]. A correlation of 95% was observed between the tweets containing the *flu* keywords and the actual national health statistics. A similar study was conducted by Corley et al. who found a high correlation between the frequency of the tweets (weekly) containing influenza keywords and the CDC[3] influenza-like-illness surveillance data [17]. Bodnar et al. compared different regression-based models for disease detection using Twitter, and discovered that the SVM regression model gave the best correlation with the actual CDC disease report [18]. Heaivilin et al. introduced Twitter as a potential source for dental surveillance and research [19]. The findings suggest that people who experience dental pain usually turn to social network to seek comfort and advice from others who also suffer from dental pain. In all such applications, systems are needed to automatically, accurately, and efficiently identify and interpret health-related content in short text "micro" messages.

Even though social media is high in noise due to the heterogeneity of the writing styles, formality, and creativity, such noise also bears undiscovered wisdom of the crowd, and hence should not be regarded as a threat, but an opportunity for discovering knowledge that can be useful in biomedical domains. Indeed literature illustrates rich research in mining biomedical and health related knowledge in social media. Paul and Dredze utilized a modified Latent Dirichlet Allocation [20] model to identify 15 ailments along with descriptions and symptoms in Twitter data [21,22]. Cameron et al. proposed a web platform *PREDOSE* (PREscription Drug abuse Online Surveillance and Epidemiology), which aims to facilitate research in prescription-related drug abuse practices using social media [23]. Greene et al. studied the quality of communication of the content in Facebook communities dedicated to diabetes. They classified each Facebook post into one of the 5 categories: Advertisements, Providing Information, Requesting Information, Support, and Irrelevant, and found that roughly two third of the information is about sharing diabetes management strategies [15]. Yang et al. proposed a method utilizing association mining and Proportional Reporting Ratios to discover the relationship between drugs and averse reactions from the user contributed content in social media [24].

This paper presents a novel machine learning based methodology that combines multi-aspect learners to make collective decisions in order to discover health-related information in the heterogeneous pool of social media. Such a system could prove useful to multiple biomedical research and applications aiming to employ the power of large-scale, realtime social media. Social media posts/comments are usually represented as short textual expressions. We formulate the problem as a text classification problem, where the objective is to correctly classify health-related content, given a large, dynamic stream of data. A message is said to be *health-related* if at least one of these two following conditions is met:

- The message indicates its author has health issues; e.g. *Fever, back pain, headache. . .ugh!*
- The message talks about someone else getting sick, or expresses health concern; e.g. *I completely understand, more than anyone! Try a warm bath too. That always helped me w/ Pauly.* & *drinking water.*

The health-related content-identification problem is transformed into the health-related short text classification, where a system is given a short text message and asked to determine whether it is health-related or not. Studies [25–27] show that traditional text classification approaches which represent a document as a "bag of words" are not well suited for processing short texts, as they do not provide sufficient word co-occurrence or shared semantics for effective similarity measures. Specifically, traditional techniques such as *N*-gram feature extraction limit the ability to recognize high-discriminative terms that include health-related keywords and/or obtain meaning from the topical semantics of the entire text. We propose and test the efficacy of ensemble methods wherein multiple base classifiers that learn different aspects of the data are used together to make collective decisions in order to enhance performance of health-related message classification.

In an effort to mitigate the limitations of existing health-related text mining methodologies, this work:

1. Proposes to use 5 heterogeneous feature types which represent different aspects of semantics for identification of health-related messages in social media. Parameter sensitivity is studied to find the best parameter configuration and base classifier for each feature type.
2. Explores the use of different ensemble methods that allow base classifiers trained with different feature types to make collective decisions.
3. Validates the proposed classification algorithms using empirical evaluation. Additionally, we strengthen the reasons for choosing the proposed features by showing how each feature type impacts the classification.

[1] http://www.patientslikeme.com/.
[2] http://www.parkinsonnet.info/.
[3] http://www.cdc.gov/flu/.

4. Evaluates the proposed classification algorithms on large scale, real world datasets, and shows that our proposed solutions do not only perform well on real-world data, but also generalize across multiple domains of social media with minimum assumption on the specific social media characteristics.

The rest of the paper is organized as follows. Section 2 provides background of the related works. Section 3 explains the characteristics of the dataset we use in our experiments. Section 4 discusses our proposed methods, including feature extraction along with analysis on parameter sensitivity and ensemble techniques in detail. Section 5 describes the evaluation of our proposed methods against the baseline on both small labeled data and large scale datasets. Section 6 concludes the paper.

2. Related works

The literature on text classification is extensive, hence we only discuss works closely related to ours.

2.1. Automatic identification of health-related information

Two approaches have been widely used to identify health-related content: keyword based and learning based methods. The former requires a dictionary containing the relevant words. A message is identified as relevant if it contains one or more keywords. Ginsberg et al. demonstrated that a regression model of influenza-like illness can be estimated using the proportion of flu-related Google search queries over the same period. They classified the query logs by detecting the presence of flu-related keywords [28]. Their method was implemented in Google Flu Trends,[4] a Google based service providing almost real-time estimates of flu activity for a number of countries around the world. Culotta claimed that Twitter data yielded better prediction on the actual flu rates than query logs, and proposed a methodology to correlate the quantity of flu-related tweets, identified by flu-related keyword detection, with the actual influenza-like-illness rates [29]. Corley et al. did a similar study on web blogs [17]. They identified flu-related blog posts using keywords *influenza* and *flu*. Yang et al. proposed to use association mining and the Proportional Reporting Ratios the mine relationship between drugs and their averse reactions in social media, which basically employ the co-occurrence frequency of the drug names and averse reactions [24]. They identified content containing the averse drug reactions by detecting the presence of the health-care keywords generated by applying Consumer Health Vocabulary (CHV) [30]. The keyword matching based approaches are simple to implement and do not consume much computing resources; however, such approaches do not only fail to capture keywords unknown to the dictionary, they also fail to deal with polysemy words.

Other widely used approaches transform identification problems into classification problems and utilize machine learning based classifiers to classify the data into classes. Traditional machine learning based approaches for text classification first trains a learner with a collection of labeled documents, then uses the trained learner to classify unlabeled documents. Learning based approaches solve the term disambiguation problems posed by the keyword matching approaches as they are able to learn some level of semantics of specific words from the surrounding contexts in which they appear. Collier and Doan proposed an algorithm for detecting disease-related tweets [5]. Specifically, the algorithm categorizes a tweet into musculoskeletal, respiratory, gastrointestinal, hemorrhagic, dermatological, or neurological related ailment. Their algorithm first (1) filters tweets that contain syndromic keywords defined in the BioCaster public health ontology [31], and then (2) classifies the filtered messages into one of the six predefined ailments using binary uni-grams as features. Their problem is similar to ours, except that they aim to identify tweets corresponding to specific ailments; while we address a broader range of messages related to health issues. For example, `'having a slight headache.'` would not fall into any ailment categories in their proposed methodology as the message only describes a general symptom (i.e. `headache`). In our work, we aim to capture such messages as well since a large collective knowledge of small-signal messages could reveal significant insights into the emerging trends [2]. Aramaki et al. proposed to use a Support Vector Machine based classifier to detect flu-related tweets [32]. The machine is trained with unigrams collected within the same proximity of the flu-related keywords. Paul and Dredze address the same problem as ours and propose a machine learning based classification algorithm used for identifying health-related tweets [21]. Uni-gram, bi-gram, and tri-gram binary word features are used to train a linear kernel SVM classifier. They further use the collected tweets to mine public health information using a LDA-like technique [22]. The parameters of the classifier are then tuned to obtain 90.04% precision and 32.0% recall, since classifiers with higher precision are preferred in their task which is to collect high quality health-related tweets. Besides using traditional binary *N*-gram features to train the classifier, which we point out later not to be sufficient and accurate enough for social media settings, their classification model was built and tuned on a small dataset of roughly 5 thousands Twitter messages. Our large scale experiments (see Section 5.7) reveal that their method does not adopt very well when being used on real-world, highly diversified data. Even though literature showed that *N*-gram features are sufficient for text classification tasks, such features fall short when dealing with document in social media domain.

2.2. Short message classification

The major differences between a short message or "microtext", and a traditional document includes the length and the formality of language. Classification algorithms that work for traditional documents may not succeed in the microtext domain due to the lower dimension and higher noise characterizing the data. This subsection explores literature on short text classification in addition to Section 2.1. Sriram et al. point out the limitation of bag-of-word strategies for tweet classification, and propose *8F* features, which primarily capture the information about authors and reply-to users [33]. While authorship is proved to be a potential source of information, our dataset (see Section 3) does not have such information available. Caragea et al. propose the system *EMERSE* for classifying and aggregating tweets and text messages about the Haiti earthquake disaster [4]. They train a SVM classifier with the combination of 4 feature sets: uni-grams, uni-grams with Relief feature selection [34], *abstractions* [35], and topic words generated by LDA [20]. Since the first 2 feature sets are *N*-gram based, they encounter similar limitations as our baseline. The other two feature sets are based on groups of terms, and would partially solve the disambiguation problem, but not the keyword recognition problem.

3. Datasets

3 Datasets are used in the experiments: a small Twitter dataset (*TwitterA*), a large Twitter dataset (*TwitterB*), and a large Facebook dataset. *TwitterA* is a labeled, almost balanced dataset and is mainly used to experiment and tuning the configurations of the classifiers. The other large datasets have natural distribution of

[4] http://www.google.org/flutrends.

the health-related messages, and are used to test the ability to generalize to real world, large scale data of the proposed methods.

3.1. TwitterA dataset

For consistency and scientific comparison, we use the same dataset as [21] which consists of 5,128 manually labeled tweets. This dataset is used for the small scale experiments employing the 10-fold cross validation protocol, and for training classifiers for the large scale experiments. Since we want to minimize the assumption about the properties of social text, all hashtags, retweets and user information are removed and only textual content is kept. Future steps of our research involve expanding the data sources to include other kinds of social media (such as Facebook, Google+, blogs, etc.), which may not have hashtags and other Twitter-like features, thus we focus on common features (such as textual information and timestamps) to develop a generic algorithm. Each tweet is a tuple of tweet ID and its textual content, and is labeled as either *positive* or *negative*. A message is *positive* if it is health related, and *negative* otherwise. The dataset contains 1832 (35.73%) positive and 3296 (64.27%) negative instances.

We note that although the size of the dataset may not completely capture the noise and lexical diversity presented in social media, the hundreds of millions tweets generated each day constrain the viability of established ground truth data of substantial proportion. Examining the literature, comparable or smaller sizes of manually labeled tweets are often used to validate the models proposed in many reputable and high-impact works such as [3,33,36,37]. Moreover, the dataset has much higher distribution of positive samples than real-world data (i.e. 35.73% vs. 1.34% in real-world Twitter data, see Section 3.2). This would allow the classifiers to learn more information about the positive class, which is of interest here.

3.2. TwitterB dataset

The *TwitterB* dataset comprises roughly 700 million public tweets in the United States during the period of 18 months from April 2011 to September 2012. These tweets were collected at random, hence representing a pseudo-uniform distribution of the overall tweets without biases to any topics. Only the tweet ID, time stamp, and textual information of each tweet are extracted. The extracted information is stored in compressed text files, yielding the total size of 25 GB. A random sample of 10,000 tweets from this dataset was manually labeled by 5 graduate students. We found 134 (1.34%) health-related messages. Unlike the *TwitterA* dataset, the *TwitterB* dataset has a natural distribution of health-related messages and is used for the large scale experiment (Section 5.7).

3.3. Facebook dataset

The *Facebook* dataset comprises 1,348,800 *Timeline* statuses and 3,541,772 associated comments of 113 participant Facebook users and their friends (a total of 60,776 Facebook users). Each participant user was asked for permission to collect their and their friends' Timeline posts. All identification was removed prior to storage. All the Facebook data will be destroyed upon acceptable progress of our research. Each Timeline status message and comment is treated as an individual message, from which the ID, timestamp, and textual information are extracted, for consistency with the other datasets. The final Facebook data contains roughly 5 millions messages, yielding 155 MB of size. A random sample of 10,000 messages was manually labeled by 5 graduate students, which reveals 107 (1.07%) health-related messages. Similar to the *TwitterB* dataset, this dataset is used for the large scale experiment (Section 5.7).

4. Methodology

Even though Twitter and Facebook data is used to verify our model, the expansion into diverse types of social media such as web blogs and Google+ provides a broader foundation for public health surveillance. The need to accommodate heterogeneous types of data means that it is important for us to design a method that easily generalizes across data sources with different properties.

We propose to combine 5 heterogeneous base classifiers, selected from different families of classification algorithms and shown to be state-of-the-art for text classification, each of which is trained with a different feature type explained in Section 4.2. For each feature type, 5 base classifiers listed in Section 4.1 are tried using 10-fold cross validation with different feature extraction parameter configurations. The base classifier and parameter configuration that yield the highest *F*-measure is chosen for ensemble experiments outlined in Section 4.3. Table 1 lists the abbreviations used in this paper for quick reference.

4.1. Base classification algorithms

On each feature type, we employ 5 classification algorithms drawn from different classification families namely:

Random Forest (RF) [38] is a tree-based ensemble classifier consisting of many decision trees. Random Forest is known for its resilient embedded feature selection algorithm, allowing it to feasibly learn from high-dimensional data such as text data. We use 100 trees for each RF classifier as suggested by [39].

Support Vector Machine (SVM) [40] is a function based classifier built upon the concept of decision planes that define decision boundaries. In our experiment we use the linear kernel SVM with $C = 1.0$. SVM has long been known for superior performance in text classification with word features [41].

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [42] is a rule-based classifier which implements a propositional rule learner. For each RIPPER classifier, we set the number of folds to 3, and the minimum weight of instances to 2.0.

Bernoulli NaiveBayes (NB) [43] is a simple probabilistic classifier implementing Bayes' theorem. NaiveBayes has been shown to perform superior in some text classification tasks such as spam filtering [44].

Multinomial NaiveBayes (MNB) [45] implements the Naive Bayes algorithm for multinomially distributed data, and is one of the two classic Naive Bayes variants used in text classification (where the data is typically represented as word vector counts). McCallum and Nigamcite [46] found Multinomial NaiveBayes to perform better than simple NaiveBayes, especially at larger vocabulary sizes.

We use LibSVM[5] implementation for SVM, and Weka[6] implementation for the other classifiers.

4.2. Feature sets

This section discusses the extraction of the 5 feature sets representing different views of the dataset.

4.2.1. N-gram features (NG)

N-gram features have been used extensively in text classification to learn word patterns in the training data. Let a document *d* be an ordered set of terms. An *N*-gram is a sequence of contiguous *N* terms in *d*. Here we represent a document with a union of its

[5] http://www.csie.ntu.edu.tw/cjlin/libsvm/.
[6] http://www.cs.waikato.ac.nz/ml/weka/

Table 1
List of abbreviations.

Type	Abbr.	Description
Classification algorithm	MNB	Multinomial NaiveBayes
	NB	Bernoulli NaiveBayes
	RF	Random Forest
	RIPPER	Repeated Incremental Pruning to Produce Error Reduction
	SVM	Support Vector Machine
Feature extraction parameter	Clean	Whether to remove punctuation and stopwords, and stem the message
	N	Max number of consecutive terms to form grams
	Stem	Stemming (whether to apply Porter's stemming algorithm to the message)
	Vocab	Vocabularies
	W	Weighting schemes (Binary, Frequency, TFIDF)
	Z	Number of topics
	C	Maximum number of terms in a compound
Ensemble technique	MS	Multi Staging
	RevMS	Reverse Multi Staging
	VOTE	Majority Voting
	WPA	Weighted Probability Averaging
Feature type	CB	Combined Features
	DC	Dictionary Based Compound Features
	NG	N-Gram Features
	ST	Sentiment Features
	TD	Topic Distribution Features

uni- to *N*-grams. Three different weighting schemes are explored: *Binary*, *Frequency*, and *TF-IDF*. Let S be the set of training documents, $V = \langle v_1, \ldots, v_M \rangle$ be the vocabulary extracted from S, t be the test document, and $F(t) = \langle f_1, \ldots, f_M \rangle$ be the feature vector of the test document t. We define the weighting schemes as follows:

$$f_i^{bin} = \begin{cases} 1 & \text{if } v_i \in t \text{ and } v_i \in V \\ 0 & \text{otherwise} \end{cases}$$

$$f_i^{freq} = TF(v_i, t)$$

$$f_i^{tfidf} = \begin{cases} \frac{TF(v_i,t)}{Max(TF(w,t):w \in t)} \cdot log \frac{|S|}{1+|s \in S: v_i \in s|} & \text{if } v_i \in t \\ 0 & \text{otherwise} \end{cases}$$

$TF(w, d)$ is the number of occurrences of term w in document d. Since social media messages do not conform with standard English, we also study how data cleaning and stemming have effects on the performance. Table 2 lists all the configuration parameters and their possible values for the NG feature extraction. Note, the features used in the baseline method proposed by Paul and Dredze [21] uses the $\langle clean = F, stem = F, N = 3, W = binary \rangle$ configuration.

4.2.2. Dictionary based compound features (DC)

As mentioned in Section 2.1, two drawbacks of *N*-gram features are (1) words with multiple meanings are treated the same (Ex. `cold` can be used in both disease or temperature contexts) and (2) important keywords are treated as normal words (Ex. `Xeroderma pigmentosum` is a disease name, but may not be identified as a discriminative feature by *N*-gram approaches since it is a rare disease and appears in only a few documents). Figueiredo et al. [47] propose compound features (c-features) for text classification. A compound of C terms is a group of C terms that occur in the same document. A compound with $C = 2$ is a generalized definition of term co-occurrence. Like NG features, we represent a document with the union of uni- to *N*-grams.

Compound features address the disambiguation problem, since they can identify different sets of term used in different scenarios. However, such features would not be able to address the keyword recognition problem as they cannot interpret the meaning of each term. Another problem of using full compound features is that the feature set can grow very large once all possible compounds are enumerated.

To overcome these challenges, we propose a feature selection strategy for the compound feature extraction, which we call Dictionary-based compound features (DC). Our DC feature extraction algorithm first generates all possible compounds from a document. Next, a compound that contains at least one term defined in the dictionary is kept. In our experiment we use 3 vocabularies: *disease*, *symptom*, and *anatomy*. We obtain such vocabularies from the Gemina project.[7] The disease and symptom vocabularies contain human disease and symptom names respectively, and are used due to the fact that there is a high chance that authors of the messages use these terms to identify their own or others' health conditions (i.e. *'I think I'm havin an **asthma** attack. . ..wtf am I tweeting?'* and *'feeling better. still have a bit of a **headache** though.'*). The anatomy vocabulary contains words used to name physical parts of a human body, and is used because the existence of body organ words may help disambiguating health-related terms (i.e. *'i'll throw pillows from my couch here. . .my **knees** are burning'*. In this example, *burning* can mean either *very hot* or *painful*. The presence of the word *knees* may help identify that *burning* actually has the latter meaning.). Table 3 lists all the configuration parameters and their possible values.

4.2.3. Topic distribution features (TD)

The intuition behind topic modeling is that an author has a set of topics in mind when writing a document. A topic is defined as a distribution of terms. The author then chooses a set of terms from the topics to compose the document. With such assumption, the whole document can be represented using a mixture of different topics. Topic modeling has also been successfully used to reduce the dimension of a document (where the number of dimensions is equal to the number of topics). Topic modeling strategies have also been applied in a variety of applications such as citation recommendation [48], document annotation [49,50], and text classification [4,51,52]. We employ the Latent Dirichlet Allocation algorithm for modeling topics in our work. We briefly describe the algorithm here for quick reference.

4.2.3.1. Latent Dirichlet Allocation. In text mining, the Latent Dirichlet Allocation (LDA) [20] is a generative model that allows a document to be represented with a mixture of topics. Past literature such as [53–55,27] demonstrates successful usage of LDA to model topics from given corpora. The basic intuition of LDA for topic modeling is that an author has a set of topics in mind when writing a document. A topic is defined as a distribution of terms. The author then chooses a set of terms from the topics to compose the document. With such assumption, the whole document can be represented using a mixture of different topics. LDA serves as a means to trace back the topics in the author's mind before the document is written. Mathematically, the LDA model is described as follows:

$$P(w_i|d) = \sum_{j=1}^{|Z|} P(w_i|z_i = j) \cdot P(z_i = j|d). \quad (1)$$

$P(w_i|d)$ is the probability of term w_i being in document d. z_i is the latent (hidden) topic. $|Z|$ is the number of all topics, which needs to be predetermined. $P(w_i|z_i = j)$ is the probability of term w_i being in topic j. $P(z_i = j|d)$ is the probability of picking a term from topic j in the document d.

Essentially, the aim of LDA model is to find $P(z|d)$, the topic distribution of document d, with each topic described by the distribution over all terms $P(w|z)$.

[7] http://gemina.igs.umaryland.edu.

Table 2
Parameters for NG feature extraction.

Param.	Description	Possible values
Clean	Whether to remove punctuation and lowercase the message	T, F
Stem	Whether to apply Porter's stemming algorithm to the message	T, F
N	Max number of consecutive terms to form grams	1, 2, 3
W	Weighting schemes	Binary, freq, tfidf

Table 3
Parameters for DC feature extraction.

Param.	Description	Possible values
Stem	Whether to apply Porter's stemming algorithm to the message	T, F
Vocab	Vocabularies used	Disease, symptom, anatomy, all
N	Max number of consecutive terms to form grams	1, 2, 3
C	Maximum number of terms in a compound	1, 2
W	Weighting schemes	Binary, freq, tfidf

After the topics are modeled, we can assign a distribution of topics to a given document using a technique called *inference* [56]. A document then can be represented by a vector of numbers, each of which represents the probability of the document belonging to a topic:

$$\text{Infer}(d, Z) = \langle z_1, z_2, \ldots, z_Q \rangle; \quad |Z| = Q, \tag{2}$$

where Z is a set of topics, d is a document, and z_i is a probability of the document d falling into topic i.

Here we use topic distribution to represent a document. Since a topic is represented by a group of weighted terms, one can think of a set of topics as a form of compound features, where the weighted terms in a topic represent the components in a compound, and hence we hypothesize that using topic distribution as features can address the term disambiguation problem. For example, the term `cold` may be the top terms in two topics; one is temperature-related, and the other sickness-related.

In our work, we model topics from the training documents using LDA algorithm implemented in MALLET,[8] a *MAchine Learning for Language Toolkit*, with 3000 maximum iterations and using Gibbs sampling. We obtain the topic distribution for each test document using the inference algorithm proposed by [56]. Table 4 lists all the configuration parameters for TD feature extraction.

4.2.4. Sentiment features (ST)

Our proposed sentiment features can be divided into two groups: physical and emotional based. The physical based ST features quantify the explicit illness by measuring frequency of health related keywords in each document. We use the same sets of vocabularies as in Section 4.2.2 for health-related keywords. The emotional based features measure the level of positive and negative emotions in the message, using the *SentiStrength* algorithm proposed by Thelwall et al. [37]. Table 5 lists all the features.

Our physical based ST features also serve as a dimension reduction of the DC features (with $C = 1$). Hence, such features have the potential to address the keyword recognition problem as they capture the frequency of highly relevant keywords. We also aim to investigate whether emotional based ST features can be discriminative as social messages are contaminated with emotions. All the configuration parameters are listed in Table 6.

[8] http://mallet.cs.umass.edu/.

Table 4
Parameters for TD feature extraction.

Param.	Description	Possible values
Clean	Whether to remove punctuation and stopwords, stem the message	T, F
Z	Number of topics	50, 100, 200, 400, 600, 800, 1000

Table 5
Features used in ST feature extraction, divided into two groups: physical and emotional based.

Grp	Feature name	Description
Phys	num_diseasewords	Number of disease terms
	ratio_num_diseasewords	Ratio of disease terms to all terms
	num_symptomwords	Number of symptom terms
	ratio_num_symptomwords	Ratio of symptom terms to all terms
	num_anatomywords	Number of anatomy terms
	ratio_num_anatomywords	Ratio of anatomy terms to all terms
	num_healthwords	Number of health-related words
	ratio_num_healthwords	Ratio of health-related words to all terms
Emo.	positive_emotion	Positive Emotional Level (1–5)
	negative_emotion	Negative Emotion Level (1–5)
	num_pos_emoticons	Num positive emoticons, e.g. `:)`, `(:]`
	num_neg_emoticons	Num negative emoticons, e.g. `:(`, `=(`

4.2.5. Combined features (CB)

Having a classifier that learns all the aspects of the data may be helpful when combined with other one-aspect classifiers. We create such an overall classifier by training a base classifier with combined features generated by merging all the four feature sets discussed above into a single feature set.

4.3. Ensemble methods

In this subsection, we explain the motivation for combining base classifiers and discuss the choices of ensemble methods.

4.3.1. Preliminary study and observations

We replicated the feature set used by Paul and Dredze [21] on the original dataset and 10-fold cross validated it with a SVM classifier, which yields precision of 76.68%, recall of 47.63%, and *F*-measure of 58.76% (we later use these classification results as a baseline). In post hoc examination we observed that many of the misclassifications had the following characteristics:

Keyword Recognition Problem. Messages containing highly discriminative health-related words such as *swine, chill, and burn* are classified as non-health related. E.g. *yep he's fine...was only a mild case of the* ***swine***.

Term Disambiguation Problem. Messages containing highly discriminative health-related words used in a non-health-related context are classified as health-related. E.g. *This is* ***sick****, it's snowing again.: – It's like i am living in Russia.*

Additionally, we trained 4 classifiers based on DC, TD, ST, and DC–TD–ST (combined) feature sets (see Section 4.2), respectively, and examined the classification results. The magnitude of overlaps between the *misses* (false positives + false negatives) produced by the classifier trained with the baseline feature set and the *hits* (true positives + true negatives) produced by the DC (7.21%), TD (9.26%), ST (10.82%), DC–TD–ST (9.95%) based classifiers as seen in Table 7

Table 6
Parameters for ST feature extraction.

Param.	Description	Possible values
Stem	Whether to apply Porter's stemming algorithm to the message	T, F
N	Max number of consecutive terms to form grams	1, 2, 3
Type	Types of features to include	Physical, emotional, both

Table 7
Overlaps between misclassifications (misses) of the baseline and correct classifications (hits) of the classifiers trained with proposed feature sets.

	FP ∩ TN (%)	FN ∩ TP (%)	Misses ∩ Hits (%)
DC	4.58	2.63	7.21
TD	6.73	4.09	10.82
ST	6.63	2.63	9.26
DC–TD–ST	5.17	4.78	9.95

suggests that the addition of these features may potentially increase overall performance of social media message classification.

4.3.2. Choices of ensemble methods

The 5 base classifiers trained with different feature types are combined using standard ensemble methods listed below:

Majority Voting (VOTE) Each classifier outputs either a 'yes' or 'no'. The final outcome is the majority vote of all the classifiers.
Weighted Probability Averaging (WPA) Each classifier is given a weight, where the sum of all weights is 1. Each classifier outputs a probability estimate of the positive class. The final output is the weighted average of all the classifiers.
Multi Staging (MS) Classifiers operate in order. If a classifier says 'yes', the final output is yes; otherwise the instance in passed to the next classifier to decide.
Reverse Multi Staging (RevMS) Similar to the *MS* technique, except that an instance is passed to the next classifier if the prior classifier says 'yes'.

For the *VOTE*, *MS*, and *RevMS* methods, each base classifier classifies an instance as *positive* if the probability estimate is equal to or greater than the probability cutoff, and *negative* otherwise. For the *WPA* method, an instance is classified as *positive* if the final probability estimate is equal to or greater than the probability cutoff, and *negative* otherwise. We use 10-fold cross validation to validate the classification performance. A validation set of 10% is held-out of each training fold for setting probability cutoff and selecting the weights for *WPA* based classifiers.

5. Experiment, results, and discussion

5.1. Training base classifiers

For each feature type, all the parameter configurations are 10-fold cross validated on the dataset *TwitterA* using the 5 different base classifiers listed in Section 4.1. The parameters and probability cutoff are tuned with the 10% validation data held-out of each training fold. In order to tune the probability cutoff, we scan through different cutoff values with an increment of 0.01, and choose the one that results in the best F1 when tested with the held-out data. Note that this operation can be cheaply carried out, since the probability score of each test instance is already pre-computed. The best combination of the parameter configuration and base classifier in terms of *F*-measure is chosen. Parameter sensitivity is also investigated. The performance of the best configuration of each feature type summarized in Table 8.

5.1.1. NG based classifier

SVM is chosen for the NG feature type with configuration $\langle clean = T, stem = T, N = 2, W = tfidf \rangle$, with *F*-measure of 68.19%. To study the parameter sensitivity of the NG feature extraction, we investigate (1) the effects of document preprocessing and (2) how different weighting schemes affect the performance (*F*-measure) of the SVM classifier. Fig. 1 shows the results as a function of the maximum size of grams (*N*). Fig. 1(a) compares the performance of the feature sets with different `clean` and `stem` parameters. According to the results, cleaning and stemming the data lead to higher quality of the feature sets. Fig. 1(b) compares the results of NG feature extraction with different weighting schemes. It is clearly seen that features with TFIDF weight outperform the other weighting schemes.

5.1.2. DC based classifier

A SVM classifier with the configuration $\langle stem = true, vocab = all, N = 1, C = 2, W = tfidf \rangle$ yields the best *F*-measure (56.47%). Fig. 2 shows the parameter sensitivity analysis (*F*-measure) as functions of the maximum size of grams (*N*) on the SVM classifier. Fig. 2(a) compares the performances when different vocabularies are used. It is evident that combining all the three vocabularies yields the best results. Note that the *symptom* vocabulary gives the best results among individual vocabulary sets, this is because a large number of sickness-related tweets only talk about symptoms (headache, stomachache, sore throat, etc.) without mentioning the causing disease names. Fig. 2(b) compares the results achieved with different weighting schemes. First point to note, the performances of all the weighting schemes decrease as *N* increases. This is because compounds with bigger grams tend to generate sparse and idiosyncratic features. Similar to the NG features, the TFIDF weighting scheme outperforms the others.

5.1.3. TD based classifier

Our results show that the configuration $\langle clean = F, Z = 200 \rangle$ with a Random Forest classifier yields the best *F*-measure (54.50%). As part of the parameter impact on the RF classifier, we vary the number of topics, and also model topics from both 'cleaned' and 'uncleaned' datasets. Fig. 3 shows that the optimum number of topics is 200. Too few topics may lead to broad topics, hence low discriminative power; whereas, too many topics can result in spurious, meaningless topics consisting of idiosyncratic word combinations. An unexpected research finding is that *uncleaned* data gives a better performance, contrasting with the analysis of the NG, DC, and ST features which agree that cleaning the data in the preprocess step helps remove noise and boost the performance.

5.1.4. ST based classifier

A RIPPER classifier with the configuration $\langle stem = T, N = 2, type = both \rangle$ yields the best *F*-measure (51.08%). Fig. 4 shows the results from varying `type` and `stem` parameters as a function of the maximum size of grams (*N*) when tested with a RIPPER classifier. From Fig. 4(a), it is interesting to see that the emotional-based features do not significantly help to increase the performance. This is because most Twitter users who tweet about their sicknesses do not always express negative feelings. Oftentimes, they make the messages sound humorous by adding positive emotions or use positive tones, e.g. `GWS ya bang:P T Oh no I'm sick! Gotta use some rest:) LOL`

5.1.5. CB based classifier

The combined features include all the previous 4 feature types generated with the chosen configurations mentioned earlier. The

Table 8
10-fold classification performance of the baseline, proposed base and ensemble classifiers, along with average training time and results from the tests of statistical significance on the dataset *TwitterA*. Pr %, Re %, and F1 % denote percentage precision, recall, and *F*-measure respectively. σ (F1) denotes the standard deviation of the *F*-measure of the 10-fold cross validation. ATT(s) denotes average training time in s.

Classifier	Pr %	Re %	F1 %	ΔF1 %	σ (F1)	ATT (s)	*p*-Value (McNemar's test)	Significant (McNemar's test, $\alpha = 0.05$)	*p*-Value (5×2 CV *t* test)	Significant (5×2 CV *t* test, $\alpha = 0.05$)
Baseline	76.68	47.63	58.76	0.00	0.029	493.6	–	–	–	–
NG	75.65	62.06	68.19	9.43	0.0292	124.1	0.16214	No	0.20255	No
DC	73.77	45.74	56.47	−2.29	0.0221	31.4	0.04331	Yes	0.02746	Yes
TD	70.48	44.43	54.50	−4.26	0.0149	23.7	0.00019	Yes	0.10096	NO
ST	55.87	47.05	51.08	−7.68	0.022	1.1	<0.00001	Yes	0.01225	Yes
CB	85.07	57.29	68.47	9.71	0.0276	1252.6	0.02361	Yes	0.00087	Yes
VOTE	77.32	65.24	70.77	12.01	0.0258	1947.77	0.1158	No	0.00062	Yes
WPA	80.45	74.52	77.37	18.61	0.029	1969.65	0.00639	Yes	0.00174	Yes
MS	56.51	91.93	69.99	11.23	0.0287	1976.42	0.00011	Yes	0.39846	No
RevMS	90.08	37.96	53.41	−5.35	0.0461	1946.23	0.02195	Yes	0.71431	No

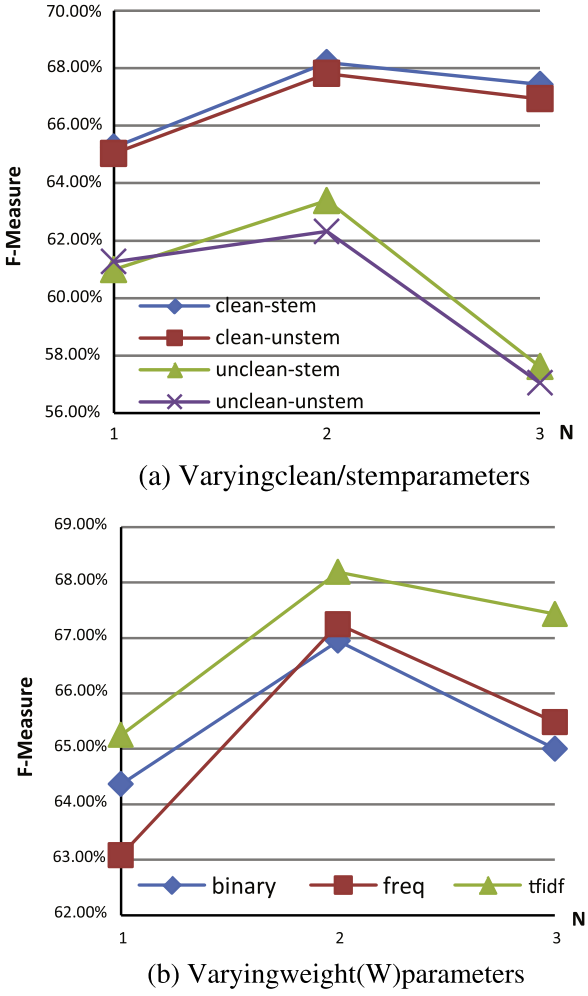


Fig. 1. Parameter comparison of NG feature extraction as the maximum size of grams (*N*).

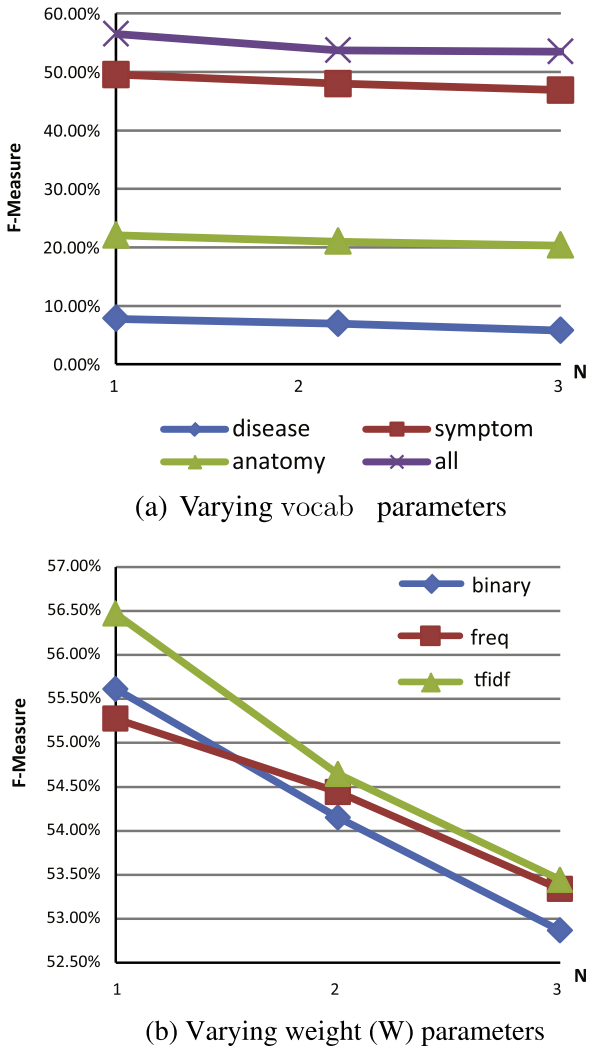


Fig. 2. Parameter comparison of DC feature extraction as the function of maximum gram size (*N*).

5 base classifiers are tried and SVM is found to perform the best with *F*-measure of 68.47%.

5.2. Small scale experiments

We evaluate each ensemble method using 10-fold cross validation on the labeled dataset *TwitterA*, using standard precision, recall, and *F*-measure (F1) as the evaluation metrics [57]. Unlike existing approaches in the literature [21] in which the quality of the retrieved data is more important than the amount, we aim to apply our algorithm in disease surveillance situations where the ability to detect non-obvious health-related messages (e.g. "*I'm not feeling good today, and prolly can't go to class.*") is also important. Hence, we treat both precision and recall as having equal importance, and *F*-measure is used to mainly compare the results from each method.

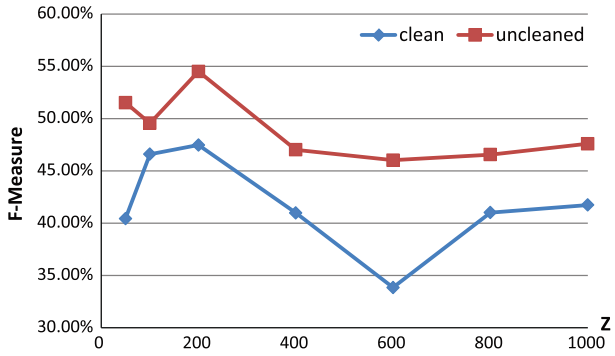
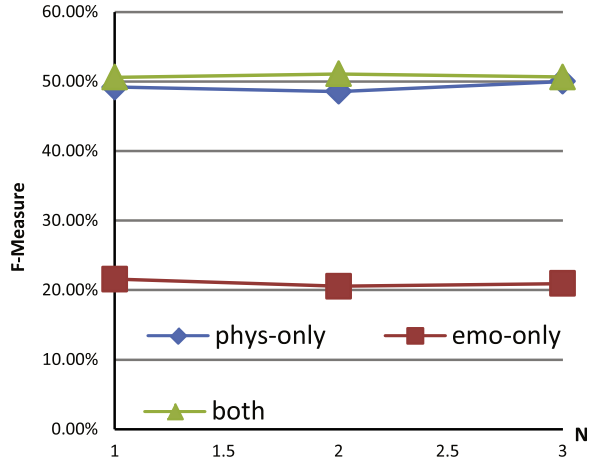


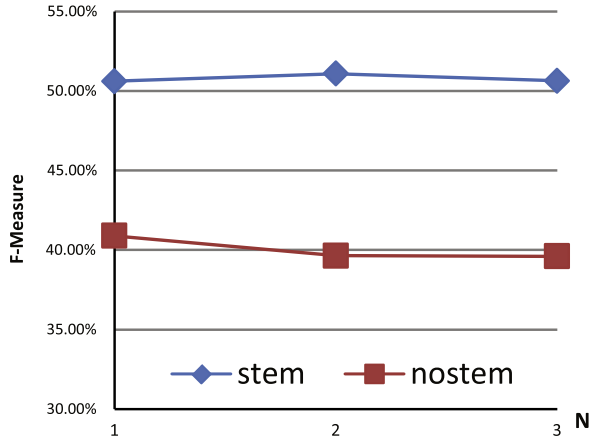
Fig. 3. Parameter comparison of TD feature extraction as the function of number of topics (Z).

The weight vectors used in the *WPA* method, the orderings of base classifiers used in the *MS* and *RevMS* methods, and the probability cutoff are tuned using 10% held-out data of the **training set** (the other 90% is used to train the base classifiers).

We compare our proposed methods with the baseline features used in related works trained with a SVM classifier tuned to achieve the best *F*-measure. Table 8 lists the results (in terms of precision, recall, *F*-measure, and *F*-measure improvement over the baseline) of each ensemble strategy, along with other base classifiers and the baseline classifier.



(a) Varying type parameter



(b) Varying stem parameter

Fig. 4. Parameter comparison of ST feature extraction as the function of maximum gram size (N).

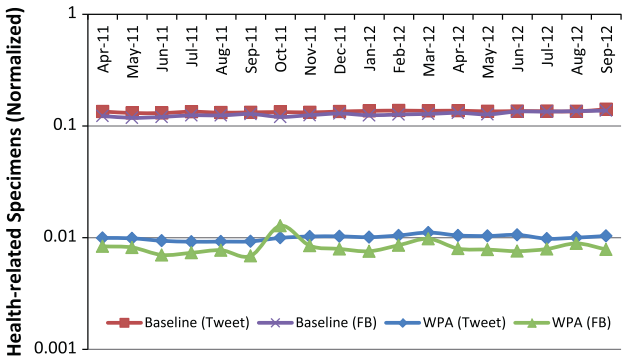


Fig. 5. A compete run of our best method (WPA) against the baseline on both the TwitterB and Facebook datasets captured during the period of 18 months from April 2011 to September 2012.

The best performance in terms of *F*-measure is yielded by the *WPA* ensemble method. This method gives some weight to all the base classifiers learning different aspects of the dataset. The *MS* method gives the best recall of 91.93%. The *RevMS* yields the best precision of 90.08%. Since we treat precision and recall as equal important, we conclude that the *WPA* ensemble method works best for our task. This might be because the *WPA* method allows all the base classifiers to make partial quantitative contribution to the final decision, hence allowing the different semantic aspects of the data to be effectively combined, as opposed to the other ensemble methods whose some base classifiers may be ignored. These results agree with a prior study of ensemble classification by Kittler et al. which found that the *sum rule* (which is a special case of the *WPA* with equal weights) outperformed other ensemble methods (i.e. multi-staging, product, maximum, median, and minimum rules) on the identity verification and the handwritten digit recognition tasks [58].

5.3. Tests of statistical significance

Two tests of statistical significance are performed to understand the statistical difference between each proposed method and the baseline: McNemar's Chi-Square Test [59] and 5×2 CV Paired t Test [60]. These two tests are chosen due to the reported low type I error by Dietterich when used to compare two supervised classification learning models [60]. Here, the null hypothesis is that each proposed model is identical to the baseline model, which is rejected if the calculated p-value is smaller than the significance level $\alpha = 0.05$.

5.3.1. McNemar's Chi-Square Test

To apply McNemar's test [61], the data S from the dataset *TwitterA* is randomly divided into a training set R (90%) and the test set T (10%). The baseline and each of the proposed models are trained using the data from R, and tested on the data from T. For each proposed algorithm f_A, the classified results are recorded in a contingency table against the baseline f_B:

Number of test instances misclassified by both f_A and f_B (n_{00})	Number of test instances misclassified by f_A, but not f_B (n_{01})
Number of test instances misclassified by f_B, but not f_A (n_{10})	Number of test instances misclassified by neither f_A nor f_B (n_{11})

Under the null hypothesis the two algorithms should have the same error rate (i.e. $n_{01} = n_{10}$). McNemar's test is based on a χ^2 test with 1 degree of freedom and is calculated as follows:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (3)$$

The above equation incorporates a *continuity correction* term (i.e. the -1 in the numerator) to account for the fact that the statistic is discrete while χ^2 distribution is continuous [60].

5.3.2. 5 × 2 CV Paired t test

McNemar's test has a drawback when dealing with small datasets: it does not measure the variability in choosing the training sets, which can significantly affect the performance of the classification models. To mitigate such an issue, Dietterich proposed the 5 × 2 cross validation paired *t* test which performs five replications of twofold cross validation. In each replication, the dataset *TwitterA* is randomly divided into two equal subsets, S_1 and S_2. For each proposed model f_A and the baseline f_B, the models are trained on each set and tested with the other set. This produces four error estimates: $p_A^{(1)}$ and $p_B^{(1)}$ (trained on S_1 and tested on S_2) and $p_A^{(2)}$ and $p_B^{(2)}$ (trained on S_2 and tested on S_1). Let $p^{(1)} = p_A^{(1)} - p_B^{(1)}$, $p^{(2)} = p_A^{(2)} - p_B^{(2)}$, and $\bar{p} = (p^{(1)} + p^{(2)})/2$, the estimated variance s^2 is defined as:

$$s^2 = (p^{(1)} - \bar{p})^2 + (p^{(2)} - \bar{p})^2 \quad (4)$$

The 5 × 2 CV $\tilde{t}$ statistic is defined as follows:

$$\tilde{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}} \quad (5)$$

where s_i^2 is the calculated s^2 of the replication *i*, and $p_1^{(1)}$ is the $p^{(1)}$ of the first replication. Under the null hypothesis, $\tilde{t}$ has approximately a *t* distribution with 5 degrees of freedom [60].

5.3.3. Analysis of statistical significance tests

The *p*-values of the McNemar's Chi-square and 5 × 2 CV paired *t* tests are reported in Table 8, along with the significance interpretations using $\alpha = 0.05$ (i.e. *YES* or *NO*). Both tests agree that the performance of *DC*, *ST*, *CB*, and *WPA* models are statistically significantly different from the baseline. It is interesting to see that the *NG* method is not reported significantly different from the baseline by both tests. This may be because both the methods rely on the *N*-gram features, which result in similar classification results.

5.4. Misclassification analysis

100 false positive and 100 false negative instances misclassified by the best proposed method (i.e. *WPA*), are randomly selected and analyzed to determine the sources of classification error. We found that the *false positive* samples can be classified into one of the three categories based on their characteristics, as listed below:

1. **The health-related keywords are presented but used in the non-health related context. (59%)** E.g.:
   - `my laptop is kinda` ***choking*** `every 2 s! gonna install UBUNTU 1.10! any tip or suggestion?`

Note that is type of error is also one of the two main weaknesses posed by the baseline. Even though error of this type is still produced by the proposed *WPA* method, the magnitude is much smaller.

2. **The message provides health related information in a sub context, but the super context is non-health related. (21%)** E.g.:
   - `I got some facebook heat for my seemingly progressive` *breast cancer* `statement. seems to me that people DO want and end to 2nd Base`
3. **The message is mislabeled. (20%)** Some health related messages are mislabeled as non-health related. This can happen due to both accidents and misunderstanding of the labellers. E.g.
   - `Rain, sick, in bed sounds good til work`

Analyzing the 100 *false negative* samples, we also found that the error can be classified into four categories:

1. **The health-related information is small, hence may produce weak signals, compared to the surrounding context. (39%)** E.g.:
   - `I'm going to have my tuition at 10.30 &` ***I am sick***`. Well, it isn't that i love studying. It's just that the $ is given. I HAVE TO GO!`

Though the word *sick* can imply that the poster is sick. However, such a word has also extensively been used in other non-health related context that the classifier may treat such a word as a *weak* signal. Especially when the health related content is among non-health related content, the signal can be impeded by the surrounding context.

2. **The message mentions health related content which can be identified from an uncommon keyword. (29%)** E.g.:
   - `KFMA Day in the Old Pueblo...might have a serious` ***sunburn*** `tomorrow. But Switchfoot is here!:)`

This problem would have been corrected by the *DC* features if the keyword is known to the vocabulary. Note that *sunburn* is a type of skin inflammation; however, the vocabulary that we use to generate the *DC* features do not contain such a word. As a result, the classifier may not be aware that *sunburn* implies health-related information.

3. **The message is mislabeled. (23%)** E.g.:
   - `Hey pregnant chick smoking in front of the burrito place, just how do you find a brand classy enough for ya?` This example message is not health related, but was labeled as *positive*.
4. **Other. (9%)** We are not able to find common characteristics among these misclassified messages.

5.5. Importance of each feature type

Our results show that the *WPA* method, wherein each base classifier is given some decision weight, yields the best performance. This section further attempts to assess the importance of each base classifier when making collaborative decisions. We analyse the results of the WPA classifier from the first fold of the 10-fold cross validation performed in Section 5.2. The best performance is yielded by the weight vector $\langle NG = 0.1, DC = 0.2, TD = 0.1, ST = 0.1, CB = 0.5\rangle$ with 74.76% precision, 68.93% recall, and 71.88% *F*-measure. The CB classifier is given most weight due to having the most extensive view of the data. The DC classifier is given a twice higher weight compared to TD and ST classifiers since it addresses both the problems posed by the baseline, while the others address only one problem.

5.6. Effect of proposed feature sets

Each of our feature set reflects a different view of the dataset–the *NG* features reflect the word patterns used in each document, the *MC* features capture the semantics of the health related terms by capturing the usage of terms appearing together in the same document, the *TD* features extract topical semantics of the document, and the *ST* features capture the sentiment semantics of document in terms of level of illness and emotional variants. According to the results in Table 8, combining all the proposed feature sets results in a better classification. This is because classifiers trained with different views of the dataset can catch the errors of the others. In this subsection, we investigate how each of our proposed feature types increases the information learned by the baseline features.

Table 9
Performance impact of each proposed feature set on the baseline feature set.

Feature set	Pre%	Rec%	*F*%	Δ*F*%
Baseline	76.68	47.63	58.76	0.00
Baseline-NG	62.96	61.32	62.13	3.37
Baseline-DC	66.96	68.74	67.84	9.08
Baseline-TD	65.44	64.26	64.85	6.09
Baseline-ST	67.41	66.05	66.72	7.96

We generate another 4 feature sets, each of which is a combination of the baseline feature set and one of our NG, DC, TD, ST feature sets. We train a SVM classifier with each of the combined feature sets, and run a 10-fold cross validation on the dataset. We compare the results with the classifier trained solely with the baseline feature set. Table 9 lists the results.

The impact (Δ F) of the NG features is not significant since the baseline and our NG features are both *N*-gram based; hence, they provide redundant information to the classifier. The DC features have the most impact on the performance, because it addresses both the drawbacks of *N*-gram features, hence allowing the classifier to learn a different perspective of the dataset. The TD features capture the topics associated with a document. However, since a topic is defined as a distribution of terms, which is similar to *N*-gram features (where a term is given a weight), the impact of TD features is not as dominant as that of the DC features. The ST features capture both health-related keywords used and emotion in a document. Since these properties are not captured in the baseline feature set, combining the ST features with the baseline allows the classifier to learn more information as expected.

Additionally, Table 10 lists the significance test results of each *non-NG base* classifiers with respect to the *NG* base classifier, using the McNemar χ^2 test outlined in Section 5.3.1. Here, the null hypothesis states that the performance of each *non-NG* base classifier is the same as the *NG* base classifier. According to Table 10, *DC*, *TD*, and *ST* base classifiers are shown to be statistically significant from the *NG* base classifier with the significance level $\alpha = 0.05$, suggesting that the proposed base classifiers learn significantly different aspects of the data that, when combined together, result in a better performance than using the *NG* feature alone. Note also that, even though each *non-NG* base classifier performs worse than or equivalent to the *NG* base classifier (according to Table 8), the *WPA* ensemble method allows these base classifiers to contribute their semantically heterogeneous knowledge to correct each other, resulting in final decisions which are more accurate than those produced by individual experts. This phenomenon also explains why the performance (in term of *F*-measure) of the *WPA* method is much better than that of the *NG* base classifier.

Note that the classification of the *CB* base classifier does not seem to be statistically significantly different from the *NG* base classifier, according to Table 10. This is consistent with the classification performance in Table 8 which reports that the classification performance (in terms of *F*-measure) of both the *NG* and *CB* base classifiers are roughly the same, while it may be intuitive that a classifier that learns all the aspects of the data should perform much better than individual experts. An explanation for this phenomenon might be the fact that the feature space of the *NG* features (64% of the combined feature space) is much larger than those of other feature types. This huge amount of *NG* features could impede the significance of other feature types when altogether learned by a base classifier. This opens a pathforward to investigate feature selection techniques, which we consider for our future work.

Table 10
Number of features and significance test results using the McNemar's method of each feature type with respect to the *NG* features.

Feature type	Num features	McNemar χ^2 score	*p*-Value (McNemar's test)	Significant McNemar's test, $\alpha = 0.05$)
NG	41,831	–	–	–
DC	23,549	6.5693	0.010375	Yes
TD	200	21.3384	0.000004	Yes
ST	14	51.4894	< 0.000001	Yes
CB	65,594	1.7349	0.187781	No

5.7. Large scale experiment

This subsection addresses three obvious questions:

1. Is smaller dataset like *TwitterA* large and diverse enough to reflect the characteristics of social media, which is full of lexical diversity and noise?
2. Are our proposed heterogeneous features able to gain insight from such a small dataset to capture the characteristics of much larger, real-world data?
3. Are our methods generalizable to other kinds of social media?

To address the above questions, we conduct another set of experiments on real-world, large scale datasets such as *TwitterB* and *Facebook* (Section 3). Each feature type is used to trained a base classifier as outlined in Section 4.1, using 90% of the data of the *TwitterA* dataset (another 10% held-out data is used to tune the parameters when combining the base classifiers). Table 11 summarized the base classifiers trained with the proposed feature types and the baseline feature, including the number of features and training time.

A random sample of 10,000 messages are drawn from each the *TwitterB* and *Facebook* datasets, and manually labeled by 5 graduate students. The sample data of the *TwitterB* dataset contains 134 (1.34%) health-related messages. The sample data of the *Facebook* dataset contains 107 (1.07%) health-related messages. It is not surprising to see a lower percentage of health-related messages in the *Facebook* dataset, since most Facebook messages are comments to existing main posts. These comments, when treated individually, may not be able to express true semantics without presented with the accompanied comments and the original posts. Hence a Facebook message may have health-related semantics (especially those comments to a health-related post), but may be classified otherwise when interpreted individually. Fig. 6 provides an example of a Facebook Timeline post and its accompanied comments. As shown in the figure, if each message is treated individually, then only the original post and comment #1 would be classified as health-related; however, when treated as a whole conversation, all the messages should be classified as health-related since they discuss the same topic (about the original poster getting the swine flu).

The baseline classifier, our base classifiers, and our proposed ensemble classifiers are used to classify these samples. Table 12 lists the results in terms of precision, recall, F1, and *F*-measure improvement over the baseline (Δ F1). The italicised numbers are the highest number in the columns.

There are four points to note:

First, it is important to note that the performance in terms of *F*-measure of the baseline classifier drops significantly (66.72% drop in precision and 41.47% drop in F1), compared to that of the 10-fold validation results in Table 8. This is because the baseline classifier is trained with a binary-based *N*-grams features on a small dataset of roughly 5000 messages. The binary features allow the classifier to take into account only the presence of terms without considering the importance of them. The obvious drawback of such scheme is that terms with high discriminative power such as *flu*,

Table 11
Summary of the base classifier, number of features, and training time (formatted as min:s) used for each proposed feature type and the baseline features.

Feature	Base classifier	# Features	Training time
Baseline	SVM	210,191	05:18
NG	SVM	67,531	03:24
DC	SVM	26,602	00:41
TD	Random Forest	200	00:21
ST	RIPPER	14	00:01
CB	SVM	94,347	21:28

Original Post	definitely has symptoms of swine flu :(
Comment 1	thats what I thought I had and missed all my exams..turns out its just like..the normal flu..dont worry about it ...
Comment 2	i'm not:D
Comment 3	Oh Gosh!! are you OK?
Comment 4	You seem rather excited so to say :P
Comment 5	what happened?
Comment 6	oh nooo...i KNEW i shudnt have talked to you yesterday :P

Fig. 6. A sample Facebook timeline post and its accompanied comments.

Table 12
Large scale classification results by our proposed methods against the baseline on a sample of 10,000 messages from each of TwitterB and Facebook datasets.

	TwitterB				Facebook			
	Pr %	Re %	F1 %	ΔF1 %	Pr %	Re %	F1 %	ΔF1 %
Baseline	9.96	65.36	17.29	0.00	8.43	61.58	14.83	0.00
NG	28.35	49.51	36.05	18.76	28.32	45.49	34.91	20.09
DC	58.32	24.83	34.83	17.54	*66.60*	31.24	42.53	27.70
TD	56.57	19.82	29.36	12.07	57.74	17.98	27.42	12.60
ST	27.45	44.61	33.99	16.70	41.60	39.18	40.35	25.53
CB	24.99	*74.28*	37.40	20.11	26.44	*70.21*	38.42	23.59
VOTE	22.50	59.46	32.65	15.36	33.05	65.42	43.91	29.09
WPA	*62.57*	64.44	*63.49*	*46.20*	60.36	63.46	*61.87*	*47.04*
MS	51.72	29.77	37.79	20.50	44.21	36.13	39.76	24.94
RevMS	30.55	19.87	24.08	6.79	36.94	34.74	35.80	20.98

cold, headache, etc. would be treated the same as common terms (e.g. *tomorrow, the, when*) and terms with low discriminative power (i.e. terms that do not imply health-related meaning such as *Xbox, iPhone, water, etc*). When testing such a classifier on a much larger and diverse data, it is expected to see a rise in recall and a drop in precision. Our *NG* features, though also based on *N*-grams, remedy both the problems by cleaning the messages (removing common stopwords and stemming terms) and utilizing TF-IDF weights to represent each term. Cleaning messages allow the classifier to ignore the common terms. Learning TF-IDF weights enable the classifier to recognize terms with highly discriminative power. Hence, the magnitude of performance reduction of the large scale performance (47.3% drop in precision, 12.55% drop in recall, and 32.14% drop in F1) of our *NG* classifier is relatively smaller compared to that of the baseline.

Second, it is worth noting that the baseline classifier tend to produce relatively high false positive rate, due to the very low precision (9.96%) and high recall rate (65.36%). To support this claim, we run the baseline classifier and our best method (*WPA*) on the whole *TwitterB* and *Facebook* datasets. The *TwitterB* data is processed on a server with a 16-core Intel Xenon E5630 (2.5 GHz) processer and 32 GB available RAM. The process was run using 40 threads (roughly 14 day's of data per thread) and was finished within 30 h. The *Facebook* data was processed on a server with an 8-core Intel Xenon E5420 (2.50 GHz) processor and 16 GB of available RAM. The process was run using 30 threads (roughly 48 day's of data per thread) and was finished in 24 h. Fig. 5 plots the normalized results (grouped by months) from April 2011 to August 2012 in log scale. According to the large scale results in Table 12, the baseline classifier tends to favor positive classes, and hence detect health-related messages at a higher proportion than our WPA method in both *TwitterB* and *Facebook* datasets. According to the results from the large scale performance evaluation shown in Table 12, our WPA methods yields comparable recall rate with that of the baseline, but much higher precision, we conclude that the higher quantity of health-related messages detected by the baseline are mostly false positives.

Third, even though the performance of all methods tend to decrease when evaluated with large scale data, our *WPA* method still yields reasonable good performance with small performance degrade (17.88% drop in precision, 10.08% drop in recall, and 13.88% drop in F1). Our *WPA* method outperforms the baseline by 46.20% in terms of *F*-measure on the *TwitterB* dataset and 47.04% on the *Facebook* dataset. We note also that, when combining base classifiers using the *WPA* method, a prominent increase in the performance is observed in both *TwitterB* and *Facebook* datasets. This advocates our assumption earlier that a proper ensemble of individual classifiers that learn different aspects of the data could improve the efficacy of the classification.

Fourth, the large scale evaluation of all the methods on both the datasets are similar. This suggests that the textual information of both social media sources is similar in nature. Hence, a classifier trained with a data source could be expected to perform reasonably equally to other social media domains as to the one it is trained with. On another hand, this also suggests that our proposed methods can easily generalize to other domains of social media.

6. Conclusions and future work

We investigate using 5 heterogeneous feature sets representing different views of the data on machine learning ensemble methods for health-related short text classification problem. We analyse the parameter sensitivity of the feature extraction algorithms in order to obtain the best possible features from each feature type. We study the mutual effects of the feature sets by combining the base classifiers, each of which is trained with a different feature type, using standard ensemble methods. We are able to outperform the baseline by 18.61% in the small scale evaluation and 46.62% on average in the large scale evaluation, using the weighted probability averaging method. Our results are very promising and reaffirm our assumption that the limitation of the *N*-gram features on the social media domain can be reduced by combining classifiers that learn different characteristics of the data. Future works could seek to improve the classification algorithm [62,63] and to employ semi-supervised methods such as the co-training technique [64] to expand the training data with unlabeled data.

Acknowledgments

We gratefully acknowledge financial support from the Penn State Center for Integrated Healthcare Delivery Systems (CIHDS), along with useful suggestions from Paronkasom Indradat and Sung Woo Kang.

References

[1] Tucker C, Kim H. Predicting emerging product design trend by mining publicly available customer review data. In: Proceedings of the 18th international conference on engineering design (ICED11), vol. 6; 2011. p. 43–52.

[2] Tuarob S, Tucker CS. Fad or here to stay: predicting product market adoption and longevity using large scale, social media data. In: Proceedings of the ASME 2013 international design engineering technical conference on computers and information in engineering conference, IDETC/CIE '13; 2013.

[3] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, WWW '10; 2010. p. 851–60.

[4] Caragea C, McNeese N, Jaiswal A, Traylor G, Kim H, Mitra P, et al. Classifying text messages for the haiti earthquake. In: Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011); 2011.

[5] Collier N, Doan S. Syndromic classification of twitter messages. CoRR abs/1110.3094.

[6] Lopes L, Zamite J, Tavares B, Couto F, Silva F, Silva M. Automated social network epidemic data collector. In: INForum informatics symposium. Lisboa; 2009.

[7] Chira P, Nugent L, Miller K, Park T, Donahue S, Soni A, et al. Living profiles: design of a health media platform for teens with special healthcare needs. J Biomed Inform 2010;43(5):S9–S12.

[8] Brennan PF, Downs S, Casper G. Project healthdesign: rethinking the power and potential of personal health records. J Biomed Inform 2010;43(5, Supplement):S3–5. http://dx.doi.org/10.1016/j.jbi.2010.09.001.

[9] Merolli M, Gray K, Martin-Sanchez F. Health outcomes and related effects of using social media in chronic disease management: a literature review and analysis of affordances. J Biomed Inform.

[10] Terry M. Twittering healthcare: social media and medicine. Telemed e-Health 2009;15(6):507–10.

[11] Kaye J, Curren L, Anderson N, Edwards K, Fullerton SM, Kanellopoulou N, et al. From patients to partners: participant-centric initiatives in biomedical research. Nat Rev Genet 2012;13(5):371–6.

[12] Hesse B, Hansen D, Finholt T, Munson S, Kellogg W, Thomas J. Social participation in health 2.0. Computer 2010;43(11):45–52. http://dx.doi.org/10.1109/MC.2010.326.

[13] Jain SH. Practicing medicine in the age of Facebook. New Engl J Med 2009;361(7):649–51. http://dx.doi.org/10.1056/NEJMp0901277. pMID: 1967532.

[14] Eijk Mvd, Faber JM, Aarts WJ, Kremer AJ, Munneke M, Bloem RB. Using online health communities to deliver patient-centered care to people with chronic conditions. J Med Internet Res 2013;15(6):e115. http://dx.doi.org/10.2196/jmir.2476. <http://www.jmir.org/2013/6/e115/>.

[15] Greene J, Choudhry N, Kilabuk E, Shrank W. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen Internal Med 2011;26(3):287–92. http://dx.doi.org/10.1007/s11606-010-1526-3. http://dx.doi.org/10.1007/s11606-010-1526-3.

[16] Culotta A. Detecting influenza outbreaks by analyzing twitter messages. CoRR abs/1007.4748.

[17] Corley C, Cook D, Mikler A, Singh K. Using web and social media for influenza surveillance. In: Arabnia HR, editor. Advances in computational biology. Advances in experimental medicine and biology, vol. 680. New York: Springer; 2010. p. 559–64.

[18] Bodnar T, Salathé M. Validating models for disease detection using twitter. In: Proceedings of the 22nd international conference on world wide web companion, WWW '13 companion, international world wide web conferences steering committee. Republic and Canton of Geneva, Switzerland; 2013. p. 699–702.

[19] Heaivilin N, Gerbert B, Page J, Gibbs J. Public health surveillance of dental pain via twitter. J Dental Res 2011;90(9):1047–51.

[20] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

[21] Paul MJ, Dredze M. A model for mining public health topics from twitter. Tech. rep.; 2011.

[22] Paul MJ, Dredze M. You are what you tweet: analyzing Twitter for public health. In: Fifth international AAAI conference on weblogs and and social media; 2011. p. 265–72.

[23] Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, et al. Predose: a semantic web platform for drug abuse epidemiology using social media. J Biomed Inform 2013(0). http://dx.doi.org/10.1016/j.jbi.2013.07.007.

[24] Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on smart health and wellbeing, SHB '12. New York, NY, USA: ACM; 2012. p. 33–40. http://dx.doi.org/10.1145/2389707.2389714.

[25] Phan X-H, Nguyen L-M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on world wide web, WWW '08; 2008. p. 91–100.

[26] Hu X, Sun N, Zhang C, Chua T-S. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM conference on information and knowledge management, CIKM '09. New York, NY, USA: ACM; 2009. p. 919–28. http://dx.doi.org/10.1145/1645953.1646071.

[27] Jin O, Liu NN, Zhao K, Yu Y, Yang Q. Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on information and knowledge management, CIKM '11. New York, NY, USA: ACM; 2011. p. 775–84. http://dx.doi.org/10.1145/2063576.2063689.

[28] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012–4. http://dx.doi.org/10.1038/nature0763.

[29] Culotta A. Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics, SOMA '10. New York, NY, USA: ACM; 2010. p. 115–22. http://dx.doi.org/10.1145/1964858.1964874.

[30] Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24–9.

[31] Collier N, Doan S, Kawazoe A, Goodwin R, Conway M, Tateno Y, et al. Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics 2008;24(24):2940–1.

[32] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the conference on empirical methods in natural language processing, EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 1568–76.

[33] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10; 2010. p. 841–2.

[34] Kira K, Rendell L. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the national conference on artificial intelligence. John Wiley & Sons Ltd; 1992. p. 129.

[35] Silvescu A, Caragea C, Honavar V. Combining super-structuring and abstraction on sequence classification. In: Proceedings of the 2009 ninth IEEE international conference on data mining, ICDM '09. Washington, DC, USA: IEEE Computer Society; 2009. p. 986–91. http://dx.doi.org/10.1109/ICDM.2009.130.

[36] Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, HLT '11, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 151–60.

[37] Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment in short strength detection informal text. J Am Soc Inform Sci Technol 2010;61(12):2544–58. http://dx.doi.org/10.1002/asi.v61:12.

[38] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[39] Khoshgoftaar TM, Golawala M, Hulse JV. An empirical study of learning from imbalanced data using random forest. In: Proceedings of the 19th IEEE international conference on tools with artificial intelligence, ICTAI '07, vol. 02; 2007. p. 310–7.

[40] Bishop CM. Pattern recognition and machine learning (information science and statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006.

[41] Joachims T. Text categorization with support vector machines: learning with many relevant features. Springer; 1998.

[42] Cohen WW. Fast effective rule induction. In: Twelfth international conference on machine learning. Morgan Kaufman; 1995. p. 115–23.

[43] John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Eleventh conference on uncertainty in artificial intelligence. San Mateo: Morgan Kaufman; 1995. p. 338–45.

[44] Androutsopoulos I, Koutsias J, Chandrinos KV, Spyropoulos CD. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM; 2000. p. 160–7.

[45] Mccallum A, Nigam K. A comparison of event models for Naive Bayes text classification. In: AAAI-98 workshop on 'Learning for Text Categorization'; 1998.

[46] McCallum A, Nigam K, et al. A comparison of event models for Naive Bayes text classification. AAAI-98 workshop on learning for text categorization, vol. 752. Citeseer; 1998. p. 41–8.

[47] Figueiredo F, Rocha L, Couto T, Salles T, Gonçalves MA, Meira W, Jr. Word co-occurrence features for text classification. Inform Syst 2011;36(5):843–58, doi:http://dx.doi.org/10.1016/j.is.2011.02.002.

[48] Kataria S, Mitra P, Bhatia S. Utilizing context in generative bayesian models for linked corpus. In: AAAI; 2010.

[49] Tuarob S, Pouchard LC, Noy N, Horsburgh JS, Palanisamy G. Onemercury: towards automatic annotation of environmental science metadata. In: Proceedings of the 2nd international workshop on linked science 2012: Tackling Big Data, LISC '12; 2012.

[50] Tuarob S, Pouchard LC, Giles CL. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries, JCDL '13; 2013.

[51] Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB. Integrating topics and syntax. Advan Neur Inform Process Syst 2005;17:537–44.

[52] Walker DD, Lund WB, Ringger EK. Evaluating models of latent document semantics in the presence of ocr errors. In: Proceedings of the 2010 conference on empirical methods in natural language processing, EMNLP '10; 2010. p. 240–50.

[53] Kataria S, Mitra P, Bhatia S. Utilizing context in generative bayesian models for linked corpus. In: AAAI'10; 2010. p. 1.

[54] Zhang X, Mitra P. Learning topical transition probabilities in click through data with regression models. In: Proceedings of the 13th international workshop on the web and databases, WebDB '10. New York, NY, USA: ACM; 2010. p. 11:1–6. http://dx.doi.org/10.1145/1859127.1859142.

[55] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet allocation for tag recommendation. In: Proceedings of the third ACM conference on recommender systems, RecSys '09. New York, NY, USA: ACM; 2009. p. 61–8. http://dx.doi.org/10.1145/1639714.1639726.
[56] Asuncion A, Welling M, Smyth P, Teh YW. On smoothing and inference for topic models. In: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, UAI '09. Arlington, VA, United States: AUAI Press; 2009. p. 27–34.
[57] Manning CD, Raghavan P, Schtze H. Introduction to information retrieval. New York, NY, USA: Cambridge University Press; 2008.
[58] Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Patt Anal Mach Intell 1998;20(3):226–39. http://dx.doi.org/10.1109/34.667881. <http://dx.doi.org/10.1109/34.667881>.
[59] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–7. http://dx.doi.org/10.1007/BF02295996. <http://dx.doi.org/10.1007/BF02295996>.
[60] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neur Comput 1998;10(7):1895–923. http://dx.doi.org/10.1162/089976698300017197. <http://dx.doi.org/10.1162/089976698300017197>.
[61] Everitt BS. The analysis of contingency tables, vol. 45. CRC Press; 1992.
[62] S. Zelikovitz, H. Hirsh, Improving short text classification using unlabeled background knowledge to assess document similarity. In: Proceedings of the seventeenth international conference on machine learning; 2000. p. 1183–90.
[63] Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. J Am Soc Inform Sci 1990;41(6):391–407.
[64] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, COLT' 98; 1998. p. 92–100.