‘Less is more’ - Mining Useful Features from Twitter User Profiles for User Classification in the Public Health Domain

Abstract

Purpose - This work studies automated user classification on Twitter in the public health domain, a task that is essential to many public health related research on social media but has not been addressed. It aims to obtain empirical knowledge on how to optimise the classifier performance on this task.
Design/methodology/approach - A sample of 3,100 Twitter users who tweeted about different health conditions were manually coded into six most common stakeholders. We propose new, simple features extracted from the short Twitter profiles of these users, and compare a large set of classification models (including state-of-the-art) that use more complex features and with different algorithms on this dataset.
Findings - We show that user classification in the public health domain is a very challenging task, as the best result we can obtain on this dataset is only 59% in terms of F1 score. Compared to state-of-the-art, our methods can obtain significantly better (10 percentage-points in F1 on a ‘best-against-best’ basis) results when using only a small set of 40 features extracted from the short Twitter user profile texts. **Originality/value** - Our work is the first to study the different types of users that engage in health related communication on social media, applicable to a broad range of health conditions rather than specific ones studied in the previous work. Our methods are implemented as open source tools, and together with data, are the first of this kind. We believe these will encourage future research to further improve this important task.

Keywords: public health, deep learning, social media, Twitter, machine learning, data science
Paper type: original research

1. Introduction

In recent years, social media platforms such as discussion forums, and social networks, have been growing rapidly as a channel for the communication and engagement of public health related matters. Among these, Twitter has become the most commonly used platform for this purpose (Thackeray et al. 2012), due to its support for real-time dissemination of information and personal opinions. Twitter is a social networking and microblogging platform where users post and interact with messages, or ‘tweets’. It enables its users to engage in effective and real-time information sharing and dialogic relationship building with each other (Park et al., 2016). It offers interactive features such as the ability to ‘follow’ users to form networks, retweet (i.e., republish and reshare), quote, like and reply to tweets, and to embed rich media including hyperlinks, multimedia, hashtags (a notion of ‘topic’) as well as symbols within tweets.

Due to the potential of Twitter to provide insight into public views and opinions related to health and the ability to retrieve data at little cost, it has become a valuable resource for research (Moorhead et al., 2013). Currently, research based on Twitter in the health domain can be generally divided into two types: one that studies health-related content shared on Twitter, and the other studies users who engage with such content.

The majority of previous work belong to the research of **content analysis**. This covers work that apply data mining to discover novel patterns that predict future events such as disease outbreak (Szomszor et al., 2010), or enhance our existing knowledge such as pharmacovigilance (Ginn et al., 2014); studies that analyse the nature (e.g., content, quantity) of information sharing concerning particular health conditions on Twitter (Thackeray et al., 2012, Tsuya et al., 2014; Rosenkrantz et al., 2016); and research that aims to understand the impact of such shared content in terms of engaging audience and growing communities (Ferguson et al., 2014; Singh and John, 2015; Brady et al., 2017; Rabarison et al., 2017).

In contrast, work on **user analysis** in the health domain is very limited. This typically involves user profiling based on demographic characteristics or interests. We argue that this is an equally important area since the identification and characterisation of different user types allow us to understand dominant or emerging topics, influential users, the composition of a community and the information exchange patterns therein. Such knowledge will allow us to better connect information seekers with providers, which will be of key interest to public health stakeholders. For example, public health agencies and healthcare providers can better target their audience for the promotion of information and services; information seekers and service users can better find credible information to fulfill their informational needs. While there exists a wealth of literature on social media user profiling in general, these are limited to either non-health context (Tinati et al., 2012; Uddin et al., 2014), or specific health related issues such as smokers and drug users (Kim et al., 2017; Kursuncu et al., 2018). Methods and findings from these studies are ad-hoc and not directly applicable to the general public health domain.

In this work, we study the empirical task of automatically classifying Twitter users that engage in health related information sharing, using natural language processing (NLP) and machine learning techniques. We refer to the different types of users as stakeholders, representing different interests and information needs. Our contributions are empirical and include: 1) the first study on the automatic user classification in the general public health domain, while previous work only tackled single health conditions where the classification schemes are non-applicable to other problems. We propose a generic classification scheme, release both our code and data to foster further research in this area; 2) a comparative analysis of the popular machine learning algorithms and features used for social media user classification on this specific task. We show that empirically, this is a very challenging task, as many well-established methods in other domains are shown to obtain only mediocre results; 3) a new method to capture useful features based on the short Twitter profile texts of different stakeholders. Compared to state-of-the-art, such features are easier to extract, and shown to be significantly more effective on this specific task. As one of our models using only 40 features has significantly outperformed the best performing state-of-the-art (10 percentage points) that uses thousands of features extracted by complex processes (e.g., topic modelling) from tweets, as well as additional corpora.

The rest of the paper is organised as follows. Section 2 presents a brief literature review. Section 3 describes our methodology in detail. This is followed by Section 4 that presents and discusses the results. Then Section 5 discusses the limitations of this work, and Section 6 concludes this paper with future research directions.

2. Background

We firstly discuss literature in the context of public health related communication on Twitter. This includes studies of both content analysis (Section 2.1) and user analysis (Section 2.2). We then review related work from a methodological point of view, to cover automated user classification on social media in general (Section 2.3). Finally, we discuss limitations in the state of the art to motivate our research (Section 2.4).

2.1 Content analysis of public health related communication on Twitter

As mentioned before, a large number of previous studies focus on understanding the content created by Twitter users on different health related issues. Among these, many applied data mining to Twitter streams to discover new knowledge or patterns for predicting future events. Examples include public health surveillance based on Twitter, by tracking and mining tweets of particular topics (e.g., H1N1) to discover trends and make predictions of disease outbreaks (Szomszor et al., 2010; Zhang et al., 2017); topic (Paul and Dredze, 2011) and opinion mining such as patients' perception of drug safety and adverts (Curtis et al., 2017); and pharmacovigilance (Ginn et al., 2014).

Another group of work analyse the nature (e.g., content, quantity) of information sharing concerning particular health conditions on Twitter, by studying variables such as the tweeting frequency, topics indicated by keywords and hashtags, and the geographic and temporal dynamics of the tweeting behaviours (Thackeray et al., 2012, Tsuya et al., 2014; Rosenkrantz et al., 2016). For example, Tsuya et al. (2016) and Rosenkrantz et al. (2016) analysed what and how cancer patients tweeted about their experience; Xu et al. (2016) and Loeb et al. (2017) examined different tweeting behaviours between breast and prostate cancers communities on Twitter.

In addition, some work looked at the impact of content sharing in terms of engaging audience and growing communities (Ferguson et al., 2014; Singh and John, 2015; Brady et al., 2017; Rabarison et al., 2017), or providing emotional support to patients (Pagoto et al., 2014; Reavley and Pilkington, 2014). For example, Reavley and Pilkington (2014) investigated tweets discussing mental health related issues and confirmed the potential of Twitter as a channel for providing effective social support to patients.

2.2 User analysis of public health related communication on Twitter

User analysis typically involves user profiling based on demographic characteristics or interests. As discussed before, we found such work in the context of public health communication very scarce. Ferguson et al. (2014) studied Twitter conversation collected during a cardiac society conference and classified the users into nine types, such as company, PhD candidates, research fellows and nursing professionals. Borgmann et al. (2016) classified users who tweeted about urologic oncology into categories such as individual, doctor, patient, spammer, and health organisation. Kim et al. (2017) classified users that tweeted about e-cigarettes into five types specific to the domain. Rabarison et al. (2017) identified individual and organisational users during a Twitter chat session focused on health and wellness in New Orleans. Kursuncu et al. (2018) classified Twitter users in the marijuana community into retail, informed agency and personal accounts.

The methods used by these studies can be broadly divided into three types. The first (Kim et al., 2017; Kursuncu et al., 2018) uses NLP and machine learning to train a model on samples labelled with the expected user types (called ‘labelled data’). The trained model can then be used to classify new data. There exists a wide range of studies of such methods in other context, which will be discussed in details below (Section 2.4). The second adopts manual analysis (Ferguson et al., 2014; Rabarison et al., 2017) which is only suitable for small datasets. And the last (Borgmann et al., 2016) uses proprietary tools such as Symplur[1], for which there is no information available on what algorithms or user information are used for classification.

2.3 Automated user classification on the social media in general

Broadening to other domains, there is a significant number of studies on automatic classification of users on social media. These studies find their origin in the research on ‘user profiling’ (Kanoje et al., 2015), often used in recommender systems. The goal is to identify specific groups of users based on certain attributes. Due to the large amount of studies in this broader context, below we only briefly cover typical studies based on Twitter.

In terms of the target classes, in the general domain, Tinati et al. (2012) and Taxidou and Fischer (2017) categorised Twitter users based on their role in information diffusion (e.g., idea starter, viewer, and amplifier). Uddin et al. (2014) classified Twitter users into six types including personal, professional, business, spammer, news feed, and marketing services. Other studies addressed the detection of automated Twitter accounts (bots) from human users (Chu et al., 2012), organisations from individual users (McCorriston et al., 2015), students from non-students (Al-Qurishi et al., 2015), users of different occupations (Preoţiuc-Pietro et al., 2015), and social classes (Filho et al., 2014). In sports, Yang et al. (2013) classified Twitter followers of sports clubs into fans and non-fans. In politics, many studied the classification of users based on gender, age, ethnicity and political orientation (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Cohen and Ruths, 2013; Preoţiuc-Pietro et al., 2017).

[1] https://www.symplur.com/. Last retrieved in September 2018

In terms of methods, they are predominantly based on NLP and supervised machine learning techniques. During this process, a user (called ‘instance’) is represented by a set of *‘features’*, which are individual measurable properties of instances in order to differentiate them. It is also assigned one or multiple predefined categories (classes, labels, or types). A set of such instances then comprise a ‘training data set’ that is used by a ‘supervised’ machine learning algorithm that examines the features observed in the training data, and generalises patterns that can represent most - if not all - instances of each class. These patterns are then used to classify similar, unseen instances in new data.

Previous methods have used a wide range of features, that can be generally divided into content- and behaviour-based. Content-based features make use of the texts generated by users, primarily their tweets collected over a period of time. Sometimes, an additional preprocess may be applied to extract potentially more representative features from content. For example, some used statistics to capture ‘class-biased’ words or hashtags from the tweets associated with particular types of users (Pennacchiotti and Popescu, 2011; Cohen and Ruths, 2013; McCorriston et al., 2015). Some used ‘topics’ or ‘clusters’ that emerge from users’ tweets (Yang et al., 2013; Preoţiuc-Pietro et al., 2015), while others also used sentiment (Pennacchiotti and Popescu, 2011) or similarities (Kim et al., 2017) captured from tweets. Behaviour features capture how users interact with content and other users, and can include their frequency of tweets, retweets, likes, network structure based on followers, etc (Rao et al., 2010; Cohen and Ruths, 2013; Filho et al., 2014). Almost every study made use of a mixture of both types of features.

The task can be dealt with any supervised machine learning algorithms. Among these, the most popular are Support Vector Machines (SVMs, Rao et al., 2010; Cohen and Ruths, 2013; Yang et al., 2013; Filho et al., 2014; Uddin et al., 2014; Preoţiuc-Pietro et al., 2015), decision trees (Pennacchiotti and Popescu, 2011; Cohen and Ruths, 2013; Kim et al., 2017), Logistic Regression (Preoţiuc-Pietro et al., 2015; Preoţiuc-Pietro et al., 2017), Naive Bayes (Yang et al., 2013; Filho et al., 2014), Latent Dirichlet Allocation (Cohen and Ruths, 2013), and Bayesian classification (Chu et al., 2012).

2.4 Limitations of state of the art

We identify two limitations that motivated this research. First and foremost, from the task point of view, there is a lack of studies on the automatic classification of users interested in public health communication on Twitter. Work such as Kim et al. (2017) and Kursuncu et al. (2018) studied population groups that are too specific to be generalisable. While work in non-health domains studied user groups that are of little interest to the health domain. We argue there is a need to analyse general user types that engage in public health communication on Twitter. As this will enable studies of the motivations and informational needs of different user types, the information flow and interactions between user types, and emerging trends and opinions of health related topics. Practically, the benefits can be many-fold. For example, the general public may be enabled to discover credible sources of information to ask questions and fulfill their informational needs. Healthcare providers may better target their services at the right audience, or better understand their followers. Public health agencies may achieve more effective health surveillance as they are supported to discover needs from better targeted population that requires service delivery.

Second, from a methodological point of view, despite a range of well-established approaches, there is a lack of evidence that they are directly transferable across domains, as we observe that the same algorithms and features perform differently on different tasks. It is unclear that empirically, what features and classification algorithms work best for a similar task in the health domain. In particular, existing methods have primarily used features extracted from the tweet texts posted by users, but rarely used their profile texts. Complex processes such as topic modelling, sentiment analysis, clustering, and similarity measures are used to extract features from tweets, but it is unclear if the benefits of these features are transferable. Practically, feature extraction from short profile texts can be much more efficient, as it requires less data collection and processing compared to tweet based features.

3. Method

We adopt the typical workflow for the text classification task. Starting with data collection and annotation (Section 3.1), we manually analyse a Twitter user dataset to derive common user types and label each user with these types. Then using this dataset, in the second and third steps, features are extracted (Section 3.2) to represent each user, and classification models are implemented (Section 3.3), trained and evaluated (Section 3.4) using these features on the labelled data. We also describe a set of baseline and state-of-the-art methods for comparative evaluation (Section 3.5).

3.1 Data collection and annotation

To create a Twitter user dataset related to health conditions, we use a sample from the dataset collected in Zhang and Ahmed (2018). The original dataset was collected by filtering tweets using 379 disease hashtags that are believed to represent different health conditions or diseases (e.g., #Colitis) over a period of one month. It consists of around 1.5 million English tweets, from which we derived over 450,000 users. We refer to this dataset as the ZA2018 dataset.

We took a random sample of 3,100 users from this dataset, then two coders reviewed the profiles of these users using their bio (i.e., their Twitter profile text), or the most recent 20 tweets in case their bios were absent. Using a grounded theory approach informed by literature review, a protocol was developed to categorise these users into six types of stakeholders (listed below). The annotation process lasted two weeks, which involved an initial phase where the coders met frequently to discuss and resolve as many discrepancies as possible. Inter annotator agreement was assessed on a subset of 100 users that were duplicated between the two coders, and we obtained a Kappa statistics score of 0.86. This measures the extent to which different annotators will classify users in the same way, and the figure can be considered to be ‘near perfect’ agreement (Viera and Garrett, 2005). This is the typical data annotation practice commonly used in machine learning research, and is considered to create reliable labelled data. It is used in, e.g., Kim et al. (2017) and Kursuncu et al. (2018) during their studies of health related conversations on Twitter. As it is infeasible to reach each individual Twitter user to obtain their true stakeholder type.

This dataset is called the ‘gold standard’, which is to be used for the training and evaluation of the classification model. The six types of stakeholders include:

- Advocate (892): individuals or organisations that mainly promote awareness of certain health conditions.
- Individual Health service Providers (IHP, 365): health professionals such as doctors, physicians, nurses, carers etc., who may offer advice and promote awareness. Their profiles are usually personal, rather than representing the organisations they may be affiliated to.
- Organisational Health service Providers (OHP, 273): organisations providing health services, such as hospitals, and companies selling products and services.
- Patient (274): people who suffer from certain health conditions themselves or share their personal experiences of some diseases.
- Researcher (333): individuals or organisations that are interested in advancing state-of-the-art. They share useful information about diseases and engage in discussion related to their area of interest and expertise.
- Other (963): a broad category including users whose tweets do not have a health-focused theme. They may have tweeted occasionally health related content, but also pay equal attention to other affairs.

Our classification scheme represents the general stakeholders in the public health domain, the identification of which could potentially benefit each other. For example, Patients may wish to seek professional advice or treatment from IHPs and OHPs; IHPs and OHPs may be interested in Researchers to keep up to date with recent findings; Advocates may want to connect with Patients for real life stories, or contact IHPs or OHPs to seek support in raising awareness.

3.2 Feature extraction

To train classification models using the labelled data, we need to represent each instance in the data (i.e., a user) using features. Following previous research, we experiment with both content and behaviour based features. Unlike previous methods where content based features are mostly extracted from a user's tweets, we propose to extract features from their short profile text. Further, we introduce a new method to extract a small set of class-biased features from the content, and we refer to this as 'dictionary' based features.

Content based features. Due to the colloquial nature of Twitter, we firstly applied a tweet normalisation tool[2] to preprocess the text. This involves, for example, spelling correction, elongated word normalisation ('yaaaay' becomes 'yay'), and word segmentation on hashtags ('#bowelcancer' becomes 'bowel cancer'). Next, the text is tokenised, with each word further lemmatised to return to its dictionary form (e.g., 'years' becomes 'year'). Stopwords such as English determiners and prepositional words were removed. The remaining words were weighted and used as content based features for the user. The specific word weighting method depends on the machine learning algorithms and will be detailed in Section 3.3.

For users without a bio (about 5% in the dataset), we collected and concatenated their most recent 20 tweets as their bio. Empirically, this resulted in better classification accuracy.

Behaviour based features. These include the number of tweets created by the user, favorited by the user, number of friends, followers, and number of 'lists' a user has. Then from the set of tweets collected for each user found in the original ZA2018 dataset, we obtained the number of new tweets and retweets, number of the user's tweets that were retweeted, or favorited by somebody else, number of URLs, mentions (of other users), media (e.g., pictures, videos), and hashtags. Also, we calculated the fraction of tweets that contained at least one hashtag, and number of different hashtags mentioned in their tweets. These are selected as a subset of those used in Kim et al. (2017), Pennacchiotti and Popescu (2011), and Uddin et al. (2014), as they are supported by the ZA2018 dataset.

Dictionary based features. We propose to extract stakeholder-specific dictionaries and use them to extract features from users bio. In Pennacchiotti and Popescu (2011), a method was introduced to extract 'prototypical words' from each class of Twitter users. The intuition is that a particular class of users may use a set of typical lexical expressions that distinguish themselves from other classes. Given n classes and each class c_i has S_i users, a 'prototypical' score of a word w for the class c_i is calculated as:

$$proto(w, c_i) = \frac{|w, S_i|}{\sum_{j=1}^{n} |w, S_j|} \qquad [1]$$

where $|w, S_i|$ is the number of times the word w is 'used' by all users for class c_i, and specifically, this means that they searched for each word in the tweets created by the users. The authors chose top k words for each class c_i and then for each user, its frequencies of using each of these words and the sum of frequencies per class are used to calculate its 'prototypical word' features. .

We introduce a new approach that is different in four ways. **First**, we use users' bio (as collected before) for each stakeholder, instead of their tweets. On the one hand, bio is arguably more informative, while a user may retweet or tweet things that are not always relevant to the stakeholder that it represents. On the other hand, collecting bio is programmatically more efficient than collecting tweets (for which a period of time must be defined and can impact on both the data quantity and quality).

Second, empirically, we notice that equation [1] often extracts words that are unique to a class but have very low frequencies, as these words are used by only a small number of users belonging to a class. Such features will have

[2] https://github.com/cbaziotis/ekphrasis, last accessed: August 2018

poor generalisation power. To address this, we compute a ‘goodness’ score that scales the prototypical score of w by its relevant frequency to the sum of frequencies of all words used by c_i, denoted by W_i :

$$goodness(w, c_i) = proto(w, c_i) \times \frac{|w, S_i|}{\sum_{w' \in W_i} |w', S_i|} \qquad [2]$$

Third, we apply the metric to two word classes only and treat them separately: nouns and verbs. On the one hand, we observe such words to be the most informative and distinctive in a user’s bio. For example, Advocates often say they want to ‘raise awareness’ and ‘provide support’; while IHPs often mention their occupations with keywords such as ‘Doctor’, ‘MD’, ‘therapist’. Other word classes are more diverse and less consistent. On the other hand, the frequency of these two word classes may not be on a comparable scale. Therefore, they should be treated separately.

Thus for each stakeholder type excluding ‘Other’ (which can include a very broad range of general users and therefore, lack consistent patterns in their vocabulary usage), we calculate the goodness scores of the nouns and verbs used by them, and select the top 100 highest ranked to create 10 dictionaries (5 stakeholders, 2 word classes each). Table 1 shows examples of these dictionaries.

Fourth, we represent a user using the above-created dictionaries in a different way. We use each dictionary to match against the bio of a user, then calculated the sum and the max of goodness scores of the matched nouns/verbs, the number of matches, and a boolean feature to indicate if at least one match is found. This gives us a total of 40 (4 feature per dictionary) stakeholder dictionary based features.

Stakeholder	**Top 5 nouns ranked by goodness score**	**Top 5 verbs ranked by goodness score**
Advocate	health, advocate, awareness, support, cancer	helping, dedicated, supporting, raising, save
IHP	nurse, health, consultant, specialist, coach	certified, eating, med, working, personalised
OHP	care, quality, provider, service, product	providing, tracking, assisted, pen, specialising
Patient	survivor, fibromyalgia, spoonie, ptsd, cfs	diagnosed, trying, fighting, hoping, know
Researcher	research, phd, researcher, university, scientist	leading, improve, developing, reviewed, researching

Table 1. Example top 5 ranked nouns and verbs extracted for each stakeholder type

3. Classification algorithms

We compare seven different classification algorithms, including five popular ones used in the state of the art, and two DNN based algorithms that have not been reported for such tasks. We describe them under ‘classic machine learning algorithms’, and ‘DNN based algorithms’ below. Further, for classic algorithms, we also study the effect of dimensionality reduction using Principal Component Analysis (PCA). This gives us a total of 12 algorithms for comparison.

Classic machine learning algorithms. We chose five classic machine learning algorithms, including a linear kernel SVM (**SVM-l**), a non-linear (Radial Basis Function) kernel SVM (**SVM-rbf**), Logistic Regression (**LR**), Stochastic Gradient Descent classifier (**SGD**), and Random Forest (**RF**). Among them, SVM-l and LR are the most popular ones, used in Uddin et al. (2014) and Preoţiuc-Pietro et al. (2017). Others are also very popular for classification tasks (Zhang et al., 2017).

When using content-based features with these algorithms, features are weighted by term frequency inverse document frequency, which assigns a higher weight to features (i.e., words) that have high frequency in a focused subset of bios. This is often referred to as the weighted ‘bag of words’ representation in the literature (Preoţiuc-Pietro et al., 2017).

Classic machine learning algorithms with PCA. A common problem with the bag of words representation is the high dimensionality and sparsity in the feature space, which may not be effective for learning. A popular approach is therefore to apply feature dimensionality reduction techniques, such as PCA to transform the feature representations before passing them to an algorithm for learning. Thus we couple PCA with each of the five algorithms before, and refer to them as PCA+?, where ? can be any one of the algorithms (e.g., PCA+SVM-l). We configure the PCA algorithm to reduce the number of features by half in all cases. This is an arbitrary decision, as our goal is not to find ways to optimize the performance of any algorithm.

DNN based algorithms. We use two DNN structures, both of which can be generalised and illustrated in Figure 1. In general, an input instance is represented using the three kinds of features described before: content, behaviour, and dictionary based. Both behaviour and dictionary based features are used as-is, while the text content is processed by a sub-DNN structure to extract 'advanced' features. We use two popular architectures for this purpose, to be detailed below. These extracted features are then concatenated with the behaviour and dictionary features into a single feature vector, that is then passed into the final softmax layer, which produces a probability distribution over the six target classes. The class with the highest probability is then chosen as the label for the input instance.

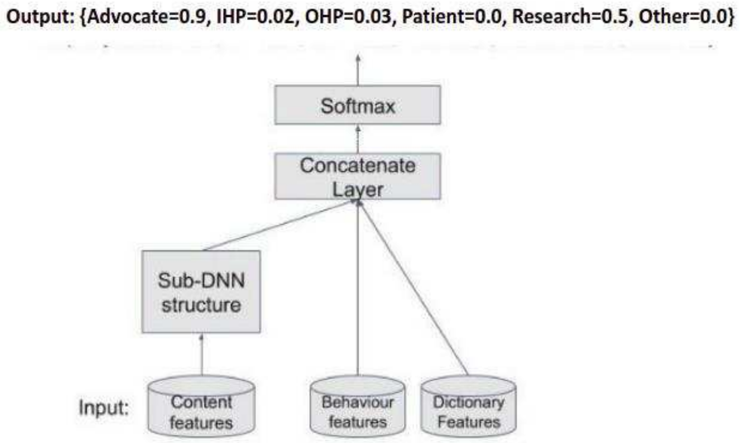

Figure 1. The generic architecture of the two DNN based classification algorithms.

We experiment with two recent DNN structures in text classification as our sub-DNN structure. The idea of these structures is to act as 'feature extractors' that prove to be effective at transforming raw, input text features into more complex, abstract features that are more effective for classification.

The first is the CNN (Convolutional Neural Networks) +'skipped' CNN structure introduced in Zhang and Luo (2018). We refer to this as **sCNN**. This starts with a 'word embedding' layer that assigns weights to each word in the input text using a fixed dimension, real-valued vector. Each dimension indicates the relative weight of the word for a 'latent' concept. These weights and latent concepts are typically pre-trained on very large text corpora. Then a dropout layer (dropout rate of 0.2) follows to 'regularise' training. The same output from the dropout layer is then passed as input to seven parallel convolutional layers, each to extract different features that are concatenated together ('joined' features). Three of these can be considered to scan a consecutive n-word sequence (a.k.a. 'window size', where n=2, 3, 4 respectively) from the input text, while the other four scans a m-word sequence (where m=3 and 4) but ignores one or two words in the middle. As an example, given a sentence containing five words 'A patient with type2 diabetes', with n=3 and m=3, the CNN layers will learn to transform sequences including 'A patient with', 'patient with type2', 'with type2 diabetes', 'A _ with', 'patient _ type2', 'with _ diabetes' into different features. Each of the parallel CNN layers uses 100 filters with a stride of 1. The joined features are then further down-sampled by a max pooling layer with a pool size of 4 and a stride of 4. All parameters of this sub-DNN structure are the same as in Zhang and Luo (2018), which we refer readers to for details.

The second is a bi-directional Long-Short Term Memory (**bi-LSTM**) network based on Lai et al. (2015). This is a type of Recurrent Neural Network (RNN) that captures long distance dependencies between words in sentences. It simulates our reading of ordered words to incrementally develop a meaning for the sentences. Bi-LSTM also starts with a word embedding layer same as that in sCNN. This is then followed by a bi-LSTM layer with 100 neurons.

For the word embedding layers in both sCNN and bi-LSTM, we use the GloVe word embedding vectors pre-trained on the Common Crawl corpus with 300 dimensions[3]. For both, we use the categorical cross entropy loss function and the Adam optimiser to train all models with an epoch of 20 using a batch size of 100.

3.4 Implementation, training and evaluation

All algorithms described above are implemented using the Scikit-Learn, Keras (2.0.9), and Theano[4] (0.9.0) libraries. Unless otherwise stated above, we used the default parameters implemented by these libraries. We share our code online to enable reproducible experiments[5]. All experiments reported in this work were conducted on a server with a maximum of 8 CPU cores and 128GB memory.

For each algorithm, we also study the impact of different types of features. For brevity, we use letters **c**, **b**, **d** to refer to content, behaviour, and dictionary based features. For each of the 12 different algorithms, we test them with 6 different feature combinations: **c**, **b**, **d**, **c+b**, **c+d**, **c+b+d**. We refer to these 72 combinations as ‘proposed models’ . As an example, SVM-l$_c$ denotes the model using the linear SVM algorithm with only content based features, while PCA+LR$_{c+b+d}$ denotes the model using PCA with Logistic Regression, with all three kinds of features.

To measure performance, we use the standard Precision, Recall, and F1 (harmonic mean of Precision and Recall) metrics used for classification tasks, calculated as below:

$$Precision = \frac{\#\,of\ true\ positives}{\#\,of\ true\ positives + \#\,of\ false\ positives} \qquad [3]$$

$$Recall = \frac{\#\,of\ true\ positives}{\#\,of\ true\ positives + \#\,of\ false\ negatives} \qquad [4]$$

$$F1 = \frac{2 \times Precision \times Recall}{(\ \ Precision + Recall)} \ [5]$$

Given a target class A, a true positive is an instance of A that is correctly classified by the model. A false positive is an instance that does not belong to A but incorrectly classified as so by the model. A false negative is an instance of A that the model fails to identify (i.e., it may be incorrectly classified as another class). Precision, Recall and F1 were calculated for each stakeholder type, and were then averaged to obtain an average score for the entire dataset (called ‘macro-average’).

There are two common approaches evaluating (i.e., testing) supervised classification methods. The first is ‘hold-out’ evaluation, where the gold standard dataset is split into two parts, one used to train a model, the other used for evaluating the trained model. The other is ‘k-fold cross-validation’, the process that splits the gold standard into k different training-evaluation pairs. The model is then trained and evaluated k times on all the pair, and then the average performance is calculated over these k runs. Practically, this gives a more reliable estimate of performance than hold-out evaluation. In this work, we use 10-fold cross-validation to evaluate all model variants. This means

[3] http://nlp.stanford.edu/data/glove.840B.300d.zip, last accessed in September 2019

[4] Scikit-Learn: https://scikit-learn.org/stable/, Kears: https://keras.io/, Theano: http://deeplearning.net/software/theano/, all of these were last accessed in September 2019

[5] Anonymised for review. Link will be shared if published.

that each model variant is trained and evaluated 10 times, each time trained on 90% of the gold standard while evaluated on the other 10%. The final model performance is the average of the figures obtained from the 10 runs.

3.5 Baseline and state-of-the-art

A major difference in our work from the previous studies is that we extract features from Twitter user profile texts instead of their tweets. To understand whether profile texts are more useful for this task, we create baseline models below. First, for each user in our dataset, we retrieve their tweets collected in the original ZA2018 dataset, and concatenate them into a single text to be called 'merged tweets'. Second, we extract content based and dictionary based features from users' merged tweets instead of their bio, giving us five alternative feature combinations: **c**, **d**, **c+b**, **c+d**, and **c+b+d**. Finally, we apply the same classification algorithms described before to these features, giving us 60 **baseline models**.

Further, we compare our methods against four state-of-the-art methods. None of the studies discussed in Section 2 released their tools or data. Therefore, we re-implement some of these methods and apply them to our dataset. However, Twitter user classification is typically domain- and task-specific, in the sense that both data and methods can be biased specifically to the task. For example, Pennacchiotti and Popescu (2011) extracted user's gender and ethnicity information, which are useful in political orientation detection but arguably, less informative for our task. They also used network based features that require collecting the Twitter network structure of users. This information is not in the ZA2018 dataset. Due to the dynamic nature of Twitter, some user accounts were deleted while for many, their networks have changed over time since the release of their dataset. Therefore, re-creating network information retrospectively will likely generate a dataset that is incomplete and skewed. In some extreme cases, certain metadata are no longer supported by the current Twitter API and therefore, cannot be obtained. For these reasons, we highlight that our re-implementations and their evaluations may not be fully representative of the methods in their original published form. Nevertheless, we have made every effort to ensure our re-implementations are as close as possible. We detail them with the adaptations below. Where features are excluded or adapted, this is because the ZA2018 dataset did not collect them or does not support the computation of them, unless otherwise stated.

- Kim2017 (Kim et al., 2017) - a method using the Gradient Boosted Regression Trees (GBRT) algorithm with 73 features, some of which are based on pairwise similarities between a user's tweets. 'Verified' is excluded for the reasons above, while 'contributors enabled', 'geo enabled', 'is translator', 'profile background tile' are no longer supported by the current Twitter API.
- Uddin2014 (Uddin et al., 2014) - a method using the linear SVM algorithm with 17 features. 'Verified', 'life time', and 'promotion score' were excluded. 'Tweet spread' is modified as the average frequency of a user's tweet being retweeted.
- Preo2015 (Preotiuc-Pietro et al., 2015) - their best performing model 'W2V-C-200' is implemented. This clusters words in a 'reference' Twitter corpus into 200 clusters, then represents each user as a weighted vector based on the distribution of words found in their tweets over these clusters (i.e., 200 features). A Gaussian Process classifier is then used for classification. This reference corpus and the parameters used for extracting these word clusters were not made available to us at the point of this work. Therefore, we used the 1.6 million public Twitter dataset for Sentiment140[6] instead, and restricted candidate words to be the most frequent 40,000 words with at least a document frequency of 5. This was the highest possible subject to our hardware limitation.
- Penn2011 (Pennacchiotti and Popescu, 2011) - 'account creation data', 'fractions of truncated tweets' and network features ('Social network: who you tweet') are excluded. For 'generic LDA' features, we used the same Sentiment140 dataset above to train 100 topics. For 'domain-specific LDA' features, we used the 1.5 million tweets from the ZA2018 dataset, excluding those belonging to the users of our dataset. For 'sentiment words', we segment the hashtags used by Zhang and Ahmed (2018) for data collection into 538

[6] http://help.sentiment140.com/for-students/, last accessed in September 2019

words as terms, and VaderSentiment[7] to classify sentiment polarity for each tweet, and associate terms found in that tweet with the polarity. GBRT is the classification algorithm. This method generates around 2,800 features.

Kim2017 studied user classification in the health domain so can be considered a very similar task to ours. Penn2011, Uddin2014, Preo2015 dealt with multinomial classification (as are us) while the majority of state-of-the-art dealt with binary classification. All methods used users' tweets to extract features. Except Uddin2014, they all processed tweets to extract complex features. Penn2011 also used an additional corpus.

4. Findings and discussions

In the following, we firstly discuss findings from the evaluation results. We then conduct an error analysis to discover the common mistakes made by the classifiers and discuss strategies to address them in the future work.

Models	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
LR$_{c}$	.56	.37	.45	.35	.37	.36	.30	.43	.36	.25	.43	.31	.41	.39	.40	.58	.56	.57	.41	.43	.41
LR$_{b}$	.35	.78	.49	.36	.01	.03	.00	.00	.00	.00	.00	.00	.29	.01	.01	.57	.65	.61	.26	.24	.19
LR$_{d}$	.58	.61	.59	.60	.51	**.55**	.64	.41	.50	.58	.33	.42	.69	.56	.62	.57	.73	.64	.61	.53	.55
LR$_{c+b}$	.56	.39	.46	.34	.36	.35	.30	.44	.36	.24	.42	.31	.40	.38	.39	.61	.58	.59	.41	.43	.41
LR$_{c+d}$	.58	.40	.47	.36	.40	.38	.30	.47	.37	.26	.48	.34	.48	.49	.49	.65	.55	.60	.44	.47	.44
LR$_{c+b+d}$	.58	.41	.48	.37	.42	.39	.30	.47	.37	.27	.49	.35	.49	.49	.49	.68	.57	.62	.45	.48	.45
RF$_{c}$	.50	.67	.57	.51	.24	.32	.51	.22	.31	.42	.14	.21	.65	.44	.53	.60	.80	.69	.53	.42	.44
RF$_{b}$	.37	.58	.45	.21	.11	.14	.28	.14	.19	.16	.05	.08	.22	.08	.12	.59	.71	.64	.30	.28	.27
RF$_{d}$	.53	.66	.59	.52	.40	.45	.54	.34	.42	.43	.24	.31	.63	.63	**.63**	.62	.68	.65	.55	.49	.51
RF$_{c+b}$	.46	.69	.55	.44	.15	.22	.53	.18	.27	.24	.03	.06	.62	.31	.42	.59	.83	**.69**	.48	.37	.37
RF$_{c+d}$	.51	.72	.59	.54	.31	.39	.56	.22	.31	.49	.14	.21	.65	.57	.61	.62	.74	.68	.56	.45	.47
RF$_{c+b+d}$	.50	.73	.59	.52	.28	.37	.57	.23	.33	.48	.11	.17	.67	.50	.57	.63	.78	.69	.56	.44	.45
SGD$_{c}$	.52	.43	.47	.47	.24	.32	.19	.33	.24	.30	.10	.15	.45	.38	.41	.54	.75	.62	.41	.37	.37
SGD$_{b}$	.32	.41	.36	.20	.14	.17	.18	.09	.12	.16	.12	.14	.14	.15	.14	.47	.51	.49	.25	.24	.24
SGD$_{d}$	.48	.53	**.60**	.44	.42	.43	.41	.47	.50	.35	.34	.34	.54	.55	.55	.53	.47	.50	.46	.46	.46
SGD$_{c+b}$	.52	.47	.50	.45	.29	.35	.18	.34	.24	.27	.09	.14	.50	.33	.39	.57	.73	.64	.41	.38	.38
SGD$_{c+d}$	.53	.53	.54	.46	.33	.38	.23	.41	.29	.34	.12	.18	.57	.43	.49	.63	.69	.66	.46	.42	.43
SGD$_{c+b+d}$	.55	.52	.54	.43	.31	.36	.23	.43	.30	.36	.17	.23	.56	.47	.51	.64	.72	.68	.46	.44	.44
SVM-l$_{c}$	.55	.34	.42	.32	.35	.33	.25	.41	.31	.23	.35	.28	.41	.38	.39	.56	.56	.56	.39	.40	.38
SVM-l$_{b}$	.39	.50	.44	.28	.14	.18	.20	.43	.28	.21	.06	.01	.21	.11	.14	.61	.60	.61	.32	.31	.29
SVM-l$_{d}$	.61	.51	.55	.51	.61	**.55**	.47	.63	**.54**	.44	.58	**.50**	.63	.62	.62	.65	.58	.61	.55	.59	**.56**
SVM-l$_{c+b}$	.55	.35	.43	.33	.35	.34	.24	.40	.30	.24	.36	.29	.40	.38	.39	.59	.59	.59	.39	.40	.39
SVM-l$_{c+d}$	.57	.39	.47	.35	.38	.36	.26	.45	.33	.25	.43	.32	.48	.46	.47	.63	.55	.59	.42	.44	.42
SVM-l$_{c+b+d}$	.57	.39	.47	.35	.39	.37	.26	.45	.33	.26	.44	.33	.48	.46	.47	.66	.57	.61	.43	.45	.43
SVM-rbf$_{c}$	.43	.32	.36	.05	.00	.01	.33	.01	.01	.00	.00	.00	.32	.02	.03	.35	.87	.50	.25	.20	.15
SVM-rbf$_{b}$	.34	.83	.48	.33	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.62	.58	.60	.22	.24	.18
SVM-rbf$_{d}$	.57	.61	.59	.58	.50	.54	.62	.38	.47	.59	.32	.41	.71	.55	.62	.59	.78	.67	.61	.52	.55
SVM-rbf$_{c+b}$	.43	.36	.39	.05	.00	.01	.33	.01	.01	.00	.00	.00	.32	.02	.03	.36	.86	.51	.25	.21	.16
SVM-rbf$_{c+d}$	.41	.63	.50	.10	.01	.01	.43	.01	.02	.00	.00	.00	.35	.02	.04	.46	.81	.59	.29	.24	.19

[7] https://github.com/cjhutto/vaderSentiment, last accessed in September 2019

SVM-rbf$_{c+b+d}$	.42	.66	.51	.10	.01	.01	.43	.01	.02	.00	.00	.00	.35	.02	.04	.48	.80	.60	.30	.25	.20

Table 2. Precision (P), Recall (R) and F1 scores obtained by the proposed models using classic machine learning algorithms, without using PCA. For each class, the highest F1 are highlighted in **bold**. For each algorithm on each class, the highest F1 are highlighted with underline.

4.1 Classification performance

Effects of different classic machine learning algorithms. Table 2 shows the Precision, Recall and F1 scores obtained by our proposed models that use the LR, RF, SGD, SVM-l and SVM-rbf algorithms with different feature sets. We summarise three patterns by comparing these algorithms in terms of F1. First, it was impossible for any model to consistently outperform others on every class, indicating the difficulty of the task. SVM-l$_{d}$ obtained the highest F1 on the IHP, OHP and Patient classes, while SGD$_{d}$ obtained the highest F1 on Advocate, and RF$_{d}$ and RF$_{c+b}$ performed best on the Research and Other classes respectively. Second, on average, SVM-l$_{d}$ is the best performer (0.56 F1), marginally beating LR$_{d}$ and SVM-rbf$_{d}$ by 1 percentage-point. Third, SVM-l models consistently outperformed SVM-rbf models regardless of the features used, and quite significantly when using feature sets other than d (dictionary based). This suggests that on this task, the linear kernel of the SVM algorithm is more robust than the non-linear, RBF kernel.

Models	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
LR$_{c}$	.49	**.53**	**.51**	**.41**	.33	**.37**	**.40**	.32	.36	**.32**	.18	.22	**.48**	.38	.41	**.61**	.54	.58	.39	**.59**	**.48**
LR$_{b}$	.35	**.80**	.49	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.55	.61	.59	.15	.24	.18
LR$_{d}$	.53	**.63**	.59	.57	.51	.54	.60	.38	.47	.55	.30	.39	.67	.52	.59	.57	.67	.62	.59	.50	.53
LR$_{c+b}$	.49	**.53**	**.51**	**.40**	.33	**.37**	**.39**	.32	.35	**.32**	.18	.22	**.45**	.38	.41	.58	**.70**	**.64**	**.43**	.41	**.42**
LR$_{c+d}$	.49	**.53**	**.51**	**.44**	**.40**	**.42**	**.31**	.35	.37	**.36**	.22	.28	**.51**	.47	.49	.62	**.71**	**.65**	**.47**	.44	**.46**
LR$_{c+b+d}$	.52	**.57**	**.55**	**.43**	.39	**.41**	**.44**	.38	**.42**	**.39**	.25	.30	**.55**	**.50**	.52	.64	**.72**	**.67**	**.50**	.47	**.48**
RF$_{c}$	.37	.58	.45	.30	.07	.12	.35	.04	.09	.12	.01	.01	.35	.08	.13	.48	.74	.58	.33	.26	.23
RF$_{b}$	.35	.49	.41	.18	.10	.13	.17	.12	.14	.06	.02	.03	.14	.07	.09	.53	.66	.60	.21	.21	.20
RF$_{d}$	.51	.57	.53	.48	**.44**	**.46**	.46	**.35**	.39	.43	**.27**	**.33**	.63	.53	.57	.59	.68	.63	.52	.47	.49
RF$_{c+b}$	.39	.60	.47	.27	.06	.11	.30	.06	.10	.00	.00	.00	.37	.06	.09	.49	.74	.59	.30	.26	.22
RF$_{c+d}$	.42	.71	.53	.38	.12	.19	**.60**	.11	.18	.24	.03	.06	.55	.28	.37	.57	.73	.64	.46	.16	.22
RF$_{c+b+d}$	.43	**.74**	.53	.41	.16	.22	.42	.07	.13	.09	.00	.02	.60	.29	.38	.50	.67	.57	.52	.39	.42
SGD$_{c}$	.44	**.56**	**.50**	**.49**	**.25**	**.33**	**.54**	.18	**.27**	**.32**	.08	.12	**.56**	.23	.32	.50	**.78**	.61	**.48**	.35	.36
SGD$_{b}$	**.33**	.35	.33	.17	**.16**	.16	**.22**	**.13**	**.16**	.09	.00	.02	.11	**.16**	.14	**.49**	**.55**	**.52**	.24	.22	.22
SGD$_{d}$	**.49**	.52	.60	.40	**.44**	.42	**.42**	.38	.46	.33	.30	.32	.52	.50	.51	.53	**.51**	**.52**	**.45**	.44	.44
SGD$_{c+b}$	.45	**.60**	**.52**	**.49**	**.30**	**.37**	**.56**	.16	**.25**	**.33**	.08	.13	**.52**	.23	.32	.52	**.77**	.63	**.48**	.36	.37
SGD$_{c+d}$	.49	**.63**	**.56**	**.51**	**.36**	**.42**	**.61**	.21	**.31**	**.41**	**.13**	**.19**	**.66**	.38	.49	.59	**.76**	**.67**	**.55**	.42	**.44**
SGD$_{c+b+d}$	.48	**.68**	**.56**	**.44**	**.33**	**.42**	**.56**	.21	.30	.36	.13	.19	**.63**	.36	.46	.71	.69	.67	.43	**.49**	.44
SVM-l$_{c}$	.52	**.51**	**.51**	**.42**	**.38**	**.40**	**.40**	.37	**.39**	**.31**	.19	.25	**.47**	**.39**	**.42**	**.57**	**.71**	**.63**	**.45**	**.43**	**.43**
SVM-l$_{b}$	.35	**.73**	**.48**	.18	.03	.04	.16	.12	.14	.00	.00	.00	.00	.00	.00	.57	.58	.58	.21	.24	.20
SVM-l$_{d}$	.58	.50	.54	.47	.61	.53	.42	.56	.48	.39	.55	.46	.64	.58	.61	.64	.52	.58	.52	.55	.53
SVM-l$_{c+b}$	.50	**.52**	**.51**	**.43**	**.38**	**.40**	**.43**	.39	.41	**.32**	.18	.22	**.47**	**.39**	**.43**	.59	**.73**	**.64**	**.46**	**.43**	**.44**
SVM-l$_{c+d}$	.55	**.55**	**.54**	**.44**	**.44**	**.44**	**.45**	.44	**.44**	**.41**	.28	**.34**	**.55**	**.54**	**.54**	**.64**	**.72**	**.67**	**.51**	**.50**	**.50**
SVM-l$_{c+b+d}$	.55	**.55**	**.54**	**.47**	**.42**	**.45**	**.46**	.45	**.45**	**.37**	.27	.31	**.53**	**.52**	**.52**	**.65**	**.74**	**.69**	**.50**	**.49**	**.50**
SVM-rbf$_{c}$	**.44**	**.38**	**.41**	**.30**	**.02**	**.0**4	**.50**	.01	.03	**.05**	.00	**.01**	**.33**	.02	**.04**	.35	.83	.50	**.33**	**.21**	**.16**
SVM-rbf$_{b}$	.34	.82	.48	**.50**	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.59	.58	.58	.24	.22	.18

SVM-rbf$_{d}$	.54	.60	.57	.54	.48	.51	.54	.33	.41	.54	.27	.36	.69	**.56**	.62	.59	.76	.66	**.58**	.50	.52
SVM-rbf$_{c+b}$	**.44**	**.47**	**.45**	**.33**	**.03**	**.05**	**.43**	.01	**.02**	**.05**	.00	**.01**	**.41**	**.03**	**.05**	**.38**	.82	**.52**	**.33**	**.22**	**.18**
SVM-rbf$_{c+d}$	**.42**	**.72**	**.53**	**.51**	**.09**	**.15**	**.64**	**.05**	**.09**	**.22**	**.02**	**.04**	**.67**	**.15**	**.25**	**.52**	.74	**.61**	**.50**	**.30**	**.28**
SVM-rbf$_{c+b+d}$	**.42**	**.75**	**.54**	**.54**	**.09**	**.17**	**.66**	**.06**	**.11**	**.27**	**.03**	**.05**	**.69**	**.16**	**.25**	**.55**	.74	**.63**	**.52**	**.31**	**.29**

Table 3. Precision (P), Recall (R) and F1 scores obtained by the proposed models using classic machine learning algorithms with PCA. If a score is higher than its corresponding model without PCA, it is highlighted in **bold**.

Effects of PCA. When PCA is used, we could not consistently obtain better performance than the highest F1 observed before. In the cases where an increase in F1 was noticed, usually c (content based) features were used. This could be due to two reasons. On the one hand, feature sets without c may already have low dimensionality. For example, as discussed before, d has only 40 different features (i.e., dimensionality of 40). Thus further reduction could have resulted in the loss of useful information for distinguishing the classes. On the other hand, content based features may be very high dimensionality due to the bag of words representation. As a result, they benefited from PCA. However, overall, we do not see strong benefits of using PCA on this task.

Effects of different DNN algorithms. Table 4 shows the Precision, Recall and F1 scores obtained by the two DNN algorithms using different feature sets. We summarise three patterns in terms of F1. First, same as before, no single model can consistently outperform others on all classes. sCNN obtained the best F1 on OHP (using c+b or c+b+d) and Other (c+b+d), while bi-LSTM obtained the best F1 on Advocate (c+b+d), IHP (c+b+d), Patient (c+d) and Researcher (c+d). Second, on average, bi-LSTM$_{c+d}$ and bi-LSTM$_{c+b+d}$ obtained the best F1 (0.59), marginally beating sCNN using the same feature sets by 1 point. Third, compared to classic algorithms, both DNN models performed much better when content based features are used (i.e., any feature sets containing c). This suggests their superior ability to capture useful features from very short text content in this task.

Model Variants	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
sCNN$_{c}$	.59	.60	.59	.54	.50	.52	.62	.48	.54	.35	.31	.33	.58	.57	.57	.73	.82	.77	.57	.55	.56
sCNN$_{b}$	.36	.76	.49	.37	.02	.04	.16	.01	.02	.07	.00	.01	.14	.01	.01	.55	.66	.60	.27	.24	.19
sCNN$_{d}$	.57	.60	.59	.58	.53	.56	.61	.47	.53	.54	.35	.43	.68	.55	.61	.59	.72	.65	.60	.54	.56
sCNN$_{c+b}$	.58	.62	.60	.54	.52	.53	.64	.53	**.58**	.36	.29	.32	.58	.56	.57	.75	.80	.77	.57	.55	.56
sCNN$_{c+d}$	.59	.63	.61	.57	.52	.54	.63	.53	.57	.40	.33	.36	.59	.61	.60	.75	.80	.77	.59	.57	.58
sCNN$_{c+b+d}$	.61	.60	.60	.55	.52	.53	.66	.52	**.58**	.38	.36	.37	.59	.60	.59	.74	.83	**.79**	.59	.57	.58
bi-LSTM$_{c}$	.61	.59	.60	.56	.53	.55	.57	.56	.56	.40	.34	.37	.60	.58	.59	.72	.80	.76	.58	.57	.57
bi-LSTM$_{b}$	.35	.76	.48	.25	.02	.04	.08	.00	.01	.00	.00	.00	.15	.01	.01	.56	.66	.61	.23	.24	.19
bi-LSTM$_{d}$	.57	.59	.58	.60	.50	.54	.63	.47	.54	.53	.32	.40	.66	.55	.60	.58	.73	.64	.59	.53	.55
bi-LSTM$_{c+b}$	.61	.62	.61	.59	.52	.55	.57	.55	.56	.44	.36	.40	.62	.59	.60	.74	.82	.78	.59	.58	.58
bi-LSTM$_{c+d}$	.61	.62	.61	.56	.53	.54	.60	.55	.57	.48	.42	**.45**	.65	.60	**.62**	.73	.79	.76	.60	.58	**.59**
bi-LSTM$_{c+b+d}$	.62	.62	**.62**	.59	.56	**.58**	.60	.52	.56	.50	.39	.43	.63	.59	.61	.71	.82	.76	.61	.58	**.59**

Table 4. Precision (P), Recall (R) and F1 scores obtained by the proposed models using DNN algorithms. For each class, the highest F1 are highlighted in **bold**. For each algorithm on each class, the highest F1 are highlighted with underline.

Effects of features. From the observations before we summarise three patterns regarding the effects of different features. First, we notice the particular effectiveness of the dictionary-based features. In terms of F1 on individual classes, almost every classic machine learning algorithm obtained their best F1 with d (See Table 2, the only exception being RF). The two DNN based algorithms also obtained competitive results when only using d features. In particular, with sCNN, d features alone contributed to the highest F1 on three classes: IHP, Patient and

Researcher. Also notice that the highest F1 scores are always obtained when d is used together with other features. Second, b (behaviour based) features are the least useful for this task, suggesting that posting patterns of Twitter users are not indicative of their stakeholder classes. Third, combining c and d features leads to the best F1 by both sCNN and bi-LSTM.

Comparison against baseline and state-of-the-art. Compared to the baseline models, overall all of our proposed models that use features extracted from profile texts outperformed their corresponding baseline model that uses features extracted from user merged tweets. To avoid presenting redundant information, we only show in Table 5 results of the baseline models using DNN algorithms. Results of other baseline models can be found in the Appendix, while we summarise the same patterns here. In terms of average F1, our proposed models obtained significantly better results than their corresponding baseline models. On a per-class basis, for classic machine learning algorithms, baseline models only performed better in very few cases. In terms of the effect of PCA on classic machine learning algorithms, we observe the same patterns described before. For DNN algorithms, all baseline models performed significantly worse in all occasions on Precision, Recall and F1,

Model Variants	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
sCNN$_c$	.39	.38	.39	.17	.16	.17	.22	.10	.13	.17	.18	.17	.22	.22	.22	.48	.57	.52	.28	.27	.27
sCNN$_d$	.38	.41	.39	.25	.07	.11	.25	.07	.10	.38	.16	.23	.33	.26	.29	.39	.64	.48	.33	.27	.27
sCNN$_{c+b}$	.39	.40	.40	.15	.08	.11	.22	.15	.18	.15	.16	.15	.24	.26	.25	.52	.61	.56	.28	.28	.27
sCNN$_{c+d}$	.38	.48	.42	.15	.12	.13	.28	.11	.16	.16	.10	.12	.23	.21	.22	.51	.59	.54	.29	.27	.27
sCNN$_{c+b+d}$	.38	.44	.41	.18	.16	.17	.23	.10	.14	.15	.16	.16	.25	.23	.24	.55	.58	.57	.29	.28	.28
bi-LSTM$_c$	.35	.42	.38	.21	.16	.18	.23	.16	.19	.23	.17	.19	.26	.24	.25	.48	.53	.50	.29	.28	.28
bi-LSTM$_d$	.38	.43	.40	.27	.08	.12	.31	.08	.13	.36	.13	.19	.35	.25	.29	.39	.63	.48	.34	.27	.27
bi-LSTM$_{c+b}$	.37	.45	.40	.26	.16	.20	.23	.18	.20	.23	.21	.22	.26	.20	.23	.52	.57	.54	.31	.29	.30
bi-LSTM$_{c+d}$	.35	.41	.38	.18	.12	.14	.26	.17	.20	.20	.20	.20	.26	.24	.25	.47	.50	.49	.29	.27	.28
bi-LSTM$_{c+b+d}$	.39	.43	.41	.22	.16	.19	.27	.19	.22	.24	.17	.20	.28	.28	.28	.50	.58	.54	.32	.30	.31

Table 5. Precision (P), Recall (R) and F1 scores obtained by the baseline models using DNN algorithms. All scores are significantly lower compared to the corresponding proposed model in Table 4.

Table 6 compares the results of the four state-of-the-art methods against our proposed models that use d features only (without PCA for classic algorithms). In terms of average F1, all of our models obtained higher F1, with six of our models (i.e., except SGD) significantly outperforming the baselines. On a 'best-against-best' basis, our sCNN$_d$ and SVN-l$_d$ models outperform Penn2011 by 10 percentage-points. On a per-class basis, our models also perform significantly better in F1 in the majority of cases.

Methods	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
Penn2011	.45	.61	.52	.58	.37	.45	.65	.32	.42	.53	.29	.38	.51	.29	.37	.54	.66	.59	.54	.42	.46
Uddin2014	.41	.28	.34	.20	.21	.21	.18	.35	.23	.24	.06	.09	.15	.02	.09	.43	.65	.52	.27	.26	.24
Preo2015	.35	.24	.29	.14	.07	.09	.08	.03	.04	.20	.06	.09	.09	.17	.12	.39	.58	.46	.20	.19	.18
Kim2017	.37	.60	.46	.25	.13	.17	.44	.25	.32	.20	.05	.09	.27	.10	.15	.60	.69	.64	.35	.30	.30
sCNN$_d$	.57	.60	.59	.58	.53	.56	.61	.47	.53	.54	.35	.43	.68	.55	.61	.59	.72	.65	.60	.54	.56
bi-LSTM$_d$	.57	.59	.58	.60	.50	.54	.63	.47	.54	.53	.32	.40	.66	.55	.60	.58	.73	.64	.59	.53	.55
LR$_d$	.58	.61	.59	.60	.51	.55	.64	.41	.50	.58	.33	.42	.69	.56	.62	.57	.73	.64	.61	.53	.55
RF$_d$	.53	.66	.59	.52	.40	.45	.54	.34	.42	.43	.24	.31	.63	.63	.63	.62	.68	.65	.55	.49	.51
SGD$_d$	.48	.53	.60	.44	.42	.43	.41	.47	.50	.35	.34	.34	.54	.55	.55	.53	.47	.50	.46	.46	.46

SVM-l$_d$	.61	.51	.55	.51	.61	.55	.47	.63	.54	.44	.58	.50	.63	.62	.62	.65	.58	.61	.55	.59	.56
SVM-rbf$_d$	.57	.61	.59	.58	.50	.54	.62	.38	.47	.59	.32	.41	.71	.55	.62	.59	.78	.67	.61	.52	.55

Table 6. Precision (P), Recall (R) and F1 scores obtained by the state-of-the-art compared to our proposed models that use d features only.

Lessons learned. We discuss three lessons learned from the observations above. First, our proposed features extracted from Twitter users' profile texts (i.e., c and d features and excluding b) work much better than features based on users' tweets. This is particularly encouraging, as compared to the baseline models and state-of-the-art, our methods have three advantages. First, it only uses short Twitter user's profile text and does not require the collection of tweets over time. Second, it does not rely on external corpora for feature extraction. Third, it uses an arguably lighter process to extract features from texts, while Kim2017, Penn2011, and Preo2015 make use of computationally expensive processes such as pairwise similarity, clustering (which in our case, even 128GB memory was insufficient), topic modelling and sentiment analysis. Overall, this suggests that, on the one hand, we can extract more effective features from Twitter users' profile texts than from their tweets. On the other hand, more features and particularly more complex features do not always translate to better performance. In the latter case, we have shown that 'less is better': when using only 40 dictionary-based features, our proposed models are able to outperform the above mentioned methods significantly. However, we reiterate that our re-implementations of state-of-the-art cannot fully represent their original form due to the inevitable reasons detailed before, and that model performance can be task- and domain-specific..

Second, among the popular machine learning algorithms used in the literature, we identify the linear SVM algorithm (SVM-l), sCNN and bi-LSTM to be the most effective. sCNN and bi-LSTM appear to be more robust than SVM-l, as their performance is less sensitive to the feature sets used, except behaviour based features that are ineffective for this task with all algorithms. In particular, they work much better with content based features. This suggests that, in situations where dictionary based feature extraction method isn't applicable (e.g., when no Twitter bio is available), sCNN or bi-LSTM will be much more reliable than other algorithms as they are able to extract highly useful features from text content only.

Finally, our results show that despite numerous methods proposed for Twitter user classification in different contexts, the task remains a very challenging one in the general public health domain, as our best average F1 is only 0.59 or 59%. This is comparatively much lower than figures reported in other domains (e.g., >70% in Uddin et al, 2014) or more specific health context (e.g., >85% in Kursuncu et al., 2018).

4.2 Error analysis

To better understand the difficulties in this task, we conduct error analysis of the classification results. We use the classification output from the bi-LSTM$_{c+d}$ model, which obtained the best average F1. We firstly create the confusion matrix, shown in Figure 2. We then select a random sample of mis-classified instances for manual inspection.

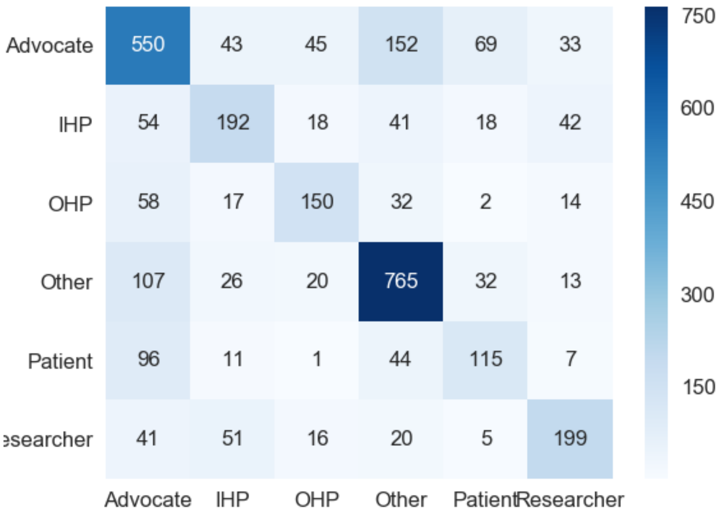

Figure 2. Confusion matrix of the output from the bi-LSTM$_{c+d}$ model. The y-axis shows the true labels; the x-axis shows the predicted labels.

Figure 2 shows that one of the most difficult tasks is to distinguish Advocate from other classes or vice versa. For example, 152 Advocate instances were misclassified as Other, while 107 Other instances were misclassified as Advocate. Patient and Advocate was the next most difficult pair to distinguish, followed by IHP, OHP and Research. The classifier also appeared to be confused very often between Patient and Other, or Research and IHP.

We took a random sample of 100 misclassified instances covering all classes, and manually coded them into different error types. We did not notice particular class-specific bias to error types. In other words, all types of errors were found in all classes. First, there are many cases of sparse (or lack of) features. For example, a very small percentage of Patients used the word ‘patient’ explicitly in their bio. Instead, some would say that they ‘struggle with’, ‘suffer from’, ‘fight’, ‘live with’, ‘battle’, ‘experience’, or ‘have’ certain conditions; and some would use expressions that imply their status, such as ‘wheelchair user’, ‘on med(ication)’ etc. This made it difficult for classifiers to generalise effective patterns. There are also cases where the bio is too short and therefore does not include useful features.

One may argue that one way to address is to bring additional text content, such as tweets created by users. However, this may not be so straightforward, as our second type of errors are due to cases where no bio is available for a user, but its most recent tweets are used instead by the classifier. In such cases, we observe that often, the content of the tweets are not necessarily indicative of the stakeholder types. Therefore, to address these two error types, a more careful design will be needed to expand a user’s Twitter bio by, possibly adding the user’s tweets in a very selective way. For example, focusing on original tweets rather than retweets, and tweets that are long and contain more contextual information (e.g., URLs, hashtags, mentions).

Third, our dataset contains a notable portion of instances that are assigned multiple stakeholder classes. This is reasonable as for example, a Patient could also be an Advocate for the health conditions that he/she is suffering from; while an OHP may be conducting Research at the same time. In fact, the bios of these instances have used words that are are indicative of their multi-class status (e.g., ‘Parkinson’s specialist, patient, caregiver, [content anonymised], & research advocate [content anonymised]’). However, our classification algorithms were designed to assign a single class to each instance. This is a rather challenging issue to address, because the true number of classes varies for each instance and is unknown a-priori. Thus a multi-label classification algorithm need to both correctly predict the true number of classes for instance and also the class labels themselves.

Further, there are a few cases of non-English profiles. These were difficult to remove during data collection as the Twitter users may have tweeted in English, and their bio could have used a combination of English and other languages.

Finally, we observe that the Other class represents a substantial part of the dataset, and some of our models have performed reasonably well on this class (e.g., 0.79 by sCNN_{c+b+d}). In theory, we may consider a two-stage process that begins with binary classification to separate it from all other classes, followed by the training of a multi-classifier on the remaining data. In practice, there is the risk of 'cascading' errors in the first stage into the second, potentially damaging the overall results. We will explore this in future work.

5. Limitations of this work

Our work is still limited in a number of ways. First, our classification scheme could have been more fine grained. For example, some advocates could be merely raising awareness, while some also provide financial or emotional support (e.g., charities offering financial support to certain patients). OHPs could be further classified based on organisational types, such as hospitals, or companies. To do so, we need to further annotate additional data in order to have a sufficient sample to train machine learning models.

Second, the performance of our classification models is far from perfect. As discussed before, a significant amount of work needs to be conducted in the future to develop other ways to capture useful features for this classification task. Possible venues could be the selective use of a user's tweets, considering a user's network, or using an ensemble of algorithms to complement each other.

6. Conclusion

Despite the significantly increasing usage of Twitter in the dissemination and engagement with public health related information, there remained a gap in research that aimed at understanding the different types of users involved in this channel of communication. This work conducted the first analysis of the common stakeholders who engage in health related communication on Twitter. Using sample data collected from Twitter, we identified six types of stakeholders commonly found in the online communication of any health conditions. A gold standard dataset was then created to develop automated classifiers using natural language processing and machine learning techniques. In particular, we proposed to use Twitter users' profile texts for feature extraction, and a new method to extract dictionary-like features from texts. These are shown to be very effective in this classification task. We believe that work in this directly will ultimately enable the research and development of solutions for many practical problems, such as better understanding the informational needs of different users, effectively connecting information seekers and providers, and identifying areas needing support from public health authorities. Our future work will explore a number of directions to address the limitations identified before for this work.

Compliance with Ethical Standards

Ethical approval: This article does not contain any studies with human participants performed by any of the authors. (nevertheless data collection from Twitter was still subject to the authors' institution's internal ethical approval)

References

1. Al-Qurishi, A., Aldrees, R., AlRubaian, M., Al-Rakhami, M., Rahman, S., Alamri, A. 2015. A new model for classifying social media users according to their behaviors. The 2nd World Symposium on Web Applications and Networking (WSWAN), Sousse, pp. 1-5. doi: 10.1109/WSWAN.2015.7209085
2. Borgmann H., Loeb S., Salem J., Thomas C., Haferkamp A., Murphy D., Tsaur, I. 2016. Activity, content, contributors, and influencers of the twitter discussion on urologic oncology. Urologic Oncology: Seminars and Original Investigations.
3. Brady R., Chapman S., Atallah S., Chand M., Mayol J., Lacy A., Wexner S. (2017) #colorectalsurgery. British Journal of Surgery. 104:1470–1476.
4. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. IEEE Trans. Dependable Secur. Comput., vol. 9, no. 6, Nov. 2012
5. Cohen, R., Ruths, D. 2013. Classifying Political Orientation on Twitter: It's not Easy! In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. 91–99.
6. Curtis J., Chen L., Higginbotham P., Nowell W., Gal-Levy R., Willig J., Safford M., Coe J., O'Hara K., Sa'adon R. (2017) Social media for arthritis-related comparative effectiveness and safety research and the impact of direct-to-consumer advertising. Arthritis Res Ther. 2017 Mar 7;19(1):48.
7. Ferguson C., Inglis S., Newton P., Cripps P., Macdonald P., Davidson P. (2014). Social media: A tool to spread information: A case study analysis of Twitter conversation at the Cardiac Society of Australia & New Zealand 61st Annual Scientific Meeting 2013. Collegian, 21(2), 89–93.
8. Filho, R., Borges, G., Almeida, J., Pappa, G. 2014. Inferring User Social Class in Online Social Networks. In Proceedings of the 8th Workshop on Social Network Mining and Analysis. ACM, New York, NY, USA
9. Ginn R., Pimpalkhute P., Nikfarjam A., Patki A., O'Connor K., Sarker A., Smith K., Gonzalez G. (2014) Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing
10. Kanoje, S., Girase, S., Mukhopadhyay, D. 2015. User Profiling Trends, Techniques and Applications, arXiv:1503.07474 [cs].
11. Kim A., Miano T., Chew R., Eggers M., Nonnemaker J. 2017. Classification of twitter users who tweet about e-cigarettes, JMIR, 26;3(3):e63.
12. Kursuncu, U., Gaur, M., Lokala, U., Illendula, A., Thirunarayan, K., Daniulaityte, R., Arpinar, I. B. (2018). "what's ur type?" contextualized classification of user types in marijuana-related communications using compositional multiview embedding. arXiv preprint arXiv:1806.06813.
13. Lai, S., Xu, L., Liu, K., Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press 2267-2273.
14. Loeb S., Gold H., Stout N., Makarov D., Weight C., Borgmann H. (2017) Tweet this: how advocacy for breast and prostate cancers stacks up on social media. BJU International (July 2017).
15. McCorriston, J., Jurgens, D., Ruths, D. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. Proceedings of the Ninth International AAAI Conference on Web and Social Media
16. Moorhead A., Hazlett D., Harrison L., Carroll J., Irwin A., Hoving C. (2013) A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. Journal of Medical Internet Research. 15(4): e85.
17. Pagoto S., Schneider K., Evans M., Waring M., Appelhans B., Busch A., Whited, M., Thind, H., Ziedonis, M. (2014). Tweeting it off: characteristics of adults who tweet about a weight loss attempt, Journal of the American Medical Informatics Association, vol. 21 (pg. 1032-1037). doi:10.1136/amiajnl-2014-00265
18. Park H., Reber, B., Chon M. (2016) Tweeting as Health Communication: Health Organizations' Use of Twitter for Health Promotion and Public Engagement, Journal of Health Communication, 21:2, 188-198
19. Paul M., Dredze M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)

20. Pennacchiotti, M., Popescu, A. (2011). A machine learning approach to Twitter user classification. 2011. Proceedings of AAAI Conference on Weblogs and Social Media.
21. Preoţiuc-Pietro, D., Lampos, V., Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (Volume 1: Long Papers). 1754-1764
22. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
23. Rabarison K., Croston M., Englar N., Bish C., Flynn S., Johnson C. (2017) Measuring Audience Engagement for Public Health Twitter Chats: Insights From #LiveFitNOLA. JMIR Public Health Surveill 2017;3(2):e34.
24. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M. 2010. Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents. ACM, New York, NY, USA, 37-44.
25. Reavley N., Pilkington P. (2014) Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. PeerJ. 2014; 2: e647
26. Rosenkrantz A., Labib A., Pysarenko K. (2016) What do patients tweet about their mammography experience? Academic Radiology 23: 1367-1371.
27. Singh K., John A. (2015). A study of tweet chats for breast cancer patients. In Proceedings of the 2015 International Conference on Social Media & Society (SMSociety '15), Anatoliy Gruzd, Jenna Jacobson, Philip Mai, and Barry Wellman (Eds.). ACM, New York, NY, USA, Article 7, 6 pages.
28. Szomszor M., Kostkova P., de Quincey E. (2009) #swineflu: Twitter predicts swine flu outbreak in 2009. In Proceedings of the 3rd International ICST Conference on Electronic Healthcare for the 21st Century (eHealth2010), Casablanca, Morocco.
29. Taxidou I., Fischer P. (2017). Structural Aspects of User Roles in Information Cascades. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1505-1509.
30. Thackeray R., Neiger B., Smith A., Van Wagenen, S. (2012). Adoption and use of social media among public health departments. BMC Public Health, 12, 242. Retrieved from http:// www.biomedcentral.com/1471-2458/12/242
31. Tinati R., Carr L., Hall W., Bentwood J. (2012). Identifying communicator roles in twitter. In Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 1161-1168.
32. Tsuya A., Sugawara Y., Tanaka A., Narimatsu H. 2014. Do cancer patients tweet? Examining the twitter use of cancer patients in Japan. J Med Internet Res. 2014;16(5):e137.
33. Uddin M., Muhammad I., Sajjad, H. 2014. Understanding types of users on Twitter. arXiv preprint arXiv:1406.1335.
34. Viera A., Garrett J. (2005). Understanding interobserver agreement: the kappa statistic. Fam Med 37:360–363.
35. Xu S., Markson C., Costello K., Xing C., Demissie K., Llanos A. (2016) Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter. JMIR Public Health Surveill 2016;2(1):e17
36. Yang, T., Lee, D., Yan, S. 2013. Steeler nation, 12th man, and boo birds: classifying Twitter user interests using time series. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, New York, NY, USA, 684-691.
37. Zhang, Z., Ahmed, W. 2018. A comparison of information sharing behaviours across 379 health conditions on Twitter. International Journal of Public Health, Volume 64, Issue 3, pp 431–440
38. Zhang, Z., Luo, L. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, vol. Pre-press, no. Pre-press, pp. 1-21.
39. Zhang, Z., Nuzzolese, A., Gentile, A. 2017. Entity deduplication on ScholarlyData. Extended Semantic Web Conference, 85-100

Appendix

Models	Advocate			IHP			OHP			Patient			Researcher			Other			**Average**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	**P**	**R**	**F1**
LR$_{c}$	.35	.25	.29	.20	.25	.23	.23	.25	.24	.21	.36	.27	.24	.33	.28	.56	.46	.51	.30	.32	.30
LR$_{d}$	.39	.45	.42	.24	.06	.10	.37	.08	.13	.44	.16	.24	.35	.24	.28	.39	.65	.49	.36	.27	.28
LR$_{c+b}$	.36	.27	.31	.20	.25	.22	.24	.25	.24	.21	.37	.27	.24	.33	.28	.61	.47	.53	.31	.32	.31
LR$_{c+d}$	.35	.22	.27	.22	.30	.25	.24	.26	.25	.20	.36	.26	.24	.35	.28	.57	.44	.50	.30	.32	.30
LR$_{c+b+d}$	.36	.23	.28	.22	.30	.25	.23	.25	.24	.20	.36	.26	.23	.35	.28	.59	.46	.52	.31	.33	.31
RF$_{c}$	.35	.59	.44	.21	.04	.07	.14	.01	.03	.20	.04	.06	.23	.05	.08	.43	.62	.51	.26	.22	.20
RF$_{d}$	.32	.45	.38	.15	.06	.09	.19	.04	.07	.26	.09	.14	.26	.15	.19	.40	.56	.47	.26	.23	.22
RF$_{c+b}$	.37	.64	.47	.27	.05	.09	.18	.03	.05	.12	.01	.03	.32	.08	.12	.50	.70	.59	.30	.25	.22
RF$_{c+d}$	.38	.62	.47	.23	.05	.09	.17	.01	.03	.23	.04	.06	.29	.07	.11	.43	.63	.51	.29	.24	.21
RF$_{c+b+d}$	.38	.65	.48	.21	.04	.07	.29	.03	.06	.29	.05	.09	.24	.05	.09	.50	.70	.58	.32	.25	.23
SGD$_{c}$	.39	.37	.38	.25	.25	.25	.26	.28	.27	.25	**.34**	**.29**	.26	.27	.27	.54	.50	.52	.33	.34	.33
SGD$_{d}$	.32	.28	.30	.21	.15	.17	.23	.16	.19	.18	.23	.20	.21	.18	.19	.35	.46	.40	.25	.24	.24
SGD$_{c+b}$	.39	.37	.38	.26	.24	.25	.27	.27	.27	.26	.34	.29	.27	.28	.28	.54	.53	.54	.33	.34	.33
SGD$_{c+d}$	.39	.36	.37	.26	.26	.26	**.28**	.29	.29	.25	**.34**	**.29**	.29	.28	.28	.53	.50	.51	.33	.34	.33
SGD$_{c+b+d}$	.39	.39	.39	.25	.24	.25	**.30**	.28	.29	.25	**.34**	**.29**	.28	.29	.28	.54	.51	.53	.34	.34	.34
SVM-l$_{c}$	.35	.22	.27	.21	.28	.24	.23	.26	.24	.20	.33	.25	.22	.34	.27	.56	.46	.51	.30	.31	.30
SVM-l$_{d}$	.45	.27	.34	.24	.16	.19	.23	.30	.26	.26	.50	.35	.28	.47	.35	.45	.41	.43	.32	.35	.32
SVM-l$_{c+b}$	.35	.22	.27	.20	.28	.23	.24	.26	.25	.20	.33	.25	.22	.33	.26	.58	.48	.53	.30	.32	.30
SVM-l$_{c+d}$	.34	.21	.26	.21	.28	.24	.23	.24	.24	.19	.34	.25	.22	.33	.26	.55	.44	.49	.29	.31	.29
SVM-l$_{c+b+d}$	.36	.22	.27	.21	.29	.25	.24	.26	.25	.19	.34	.25	.22	.34	.27	.59	.47	.52	.30	.32	.30
SVM-rbf$_{c}$	.39	.35	.36	**.33**	.01	**.02**	.00	.00	.00	**.43**	**.01**	**.02**	.00	.00	.00	.35	.83	.50	.25	.20	.15
SVM-rbf$_{d}$	.34	.51	.41	.00	.00	.00	.00	.00	.00	.29	.01	.01	**.38**	**.11**	**.16**	.38	.66	.48	**.23**	**.21**	.18
SVM-rbf$_{c+b}$	.38	.36	.37	**.33**	**.01**	**.02**	.00	.00	.00	**.43**	**.01**	**.02**	.00	.00	.00	.36	.83	.50	.25	.20	.15
SVM-rbf$_{c+d}$	.38	.38	.38	**.33**	**.01**	**.02**	.00	.00	.00	**.43**	**.01**	**.02**	.00	.00	.00	.36	.81	.50	.25	.20	.15
SVM-rbf$_{c+b+d}$	.37	.39	.38	**.33**	**.01**	**.02**	.00	.00	.00	**.43**	**.01**	**.02**	.00	.00	.00	.36	.81	.50	.25	.20	.15

Table A1. Precision (P), Recall (R) and F1 scores obtained by the baseline models using classic machine learning algorithms, without PCA. When a score is higher than its corresponding proposed model in Table 2, this is highlighted in **bold**.

Models	Advocate			IHP			OHP			Patient			Research			Other			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LR$_{c}$	**.38**	**.43**	**.40**	**.24**	.20	.22	**.28**	**.29**	**.29**	**.24**	.20	.22	**.26**	.20	.23	.51	**.53**	**.52**	**.32**	.31	**.31**
LR$_{d}$	.38	.41	.39	**.27**	.04	.07	.27	.04	.07	.39	.13	.19	.35	.20	.25	.37	**.68**	**.48**	.34	.25	.24
LR$_{c+b}$	**.40**	**.46**	**.43**	**.24**	.21	.23	.27	.26	.27	**.26**	.22	.24	.24	.18	.21	.54	**.56**	**.55**	**.32**	.32	**.32**
LR$_{c+d}$	**.39**	**.44**	**.42**	**.24**	.22	.23	**.27**	**.29**	**.28**	**.26**	.20	.22	**.25**	.20	.22	.53	**.54**	**.54**	**.32**	.31	**.32**
LR$_{c+b+d}$	**.40**	**.46**	**.43**	**.26**	.22	.24	**.29**	**.30**	**.29**	**.26**	.22	.24	**.25**	.19	.22	.53	**.54**	**.54**	**.33**	.32	**.33**
RF$_{c}$	.31	.45	.37	.19	**.06**	**.09**	**.19**	**.03**	**.05**	.20	.04	**.07**	.19	.04	.06	.38	.61	.47	.25	.21	.19

RF$_{d}$	**.34**	**.49**	**.40**	.15	**.08**	**.11**	.16	**.06**	**.09**	**.27**	.09	.14	**.29**	.15	**.20**	.39	.52	.45	**.27**	.23	**.23**
RF$_{c+b}$	.34	.49	.40	.18	.05	.08	.14	.01	.03	**.16**	**.03**	**.05**	.22	.05	.08	.41	.65	.50	.24	.22	.19
RF$_{c+d}$	.34	.48	.40	.17	.04	.07	.16	**.02**	**.04**	.26	.06	.10	.18	.05	.07	.41	**.65**	.50	.25	.22	.20
RF$_{c+b+d}$	.32	.47	.38	.17	**.05**	.07	.14	.02	.04	.16	.03	.05	.23	.05	.09	.40	.63	.49	.23	.21	.19
SGD$_{c}$	.38	.40	.39	**.27**	.22	.24	**.30**	.23	.26	**.29**	.22	.25	**.30**	.19	.24	.46	**.58**	.51	.33	.31	.31
SGD$_{d}$	.32	**.32**	**.32**	.13	.12	.13	.12	**.19**	.15	.18	.15	.16	**.24**	.18	**.21**	.35	.34	.34	.22	.22	.22
SGD$_{c+b}$	**.40**	**.42**	**.41**	**.27**	.21	.24	**.28**	.25	.27	**.30**	.27	.29	.25	.18	.21	.50	**.60**	.54	.33	.32	.32
SGD$_{c+d}$	.38	**.41**	**.40**	.24	.18	.20	**.33**	.24	.28	**.29**	.23	.26	**.31**	.21	.25	.47	**.60**	**.53**	.34	.31	.32
SGD$_{c+b+d}$	.39	**.40**	**.40**	**.28**	.22	.24	.30	.23	.26	**.31**	.26	.28	**.32**	.23	.27	**.48**	.62	.54	**.35**	.33	.33
SVM-l$_{c}$	**.42**	**.41**	**.41**	**.22**	.22	.22	**.26**	**.32**	**.29**	**.25**	.26	**.26**	**.25**	.24	.24	.53	**.52**	**.53**	**.32**	**.33**	**.32**
SVM-l$_{d}$	**.46**	.25	.33	.24	.15	.18	.22	.25	.23	.25	**.53**	.34	.26	**.49**	.34	.43	.38	.40	.31	.34	.30
SVM-l$_{c+b}$	**.42**	**.40**	**.41**	**.25**	.25	**.25**	**.26**	**.30**	**.28**	**.26**	.28	**.27**	**.26**	.25	.25	.57	**.56**	**.56**	**.34**	**.34**	**.34**
SVM-l$_{c+d}$	**.41**	**.41**	**.41**	**.24**	.23	.23	**.27**	**.33**	**.29**	**.27**	.27	**.27**	**.25**	.24	.25	.54	**.53**	**.53**	**.33**	**.33**	**.33**
SVM-l$_{c+b+d}$	**.43**	**.41**	**.42**	**.23**	.23	.23	**.26**	**.30**	**.27**	**.27**	.28	**.27**	**.28**	.26	.27	.55	**.55**	**.55**	**.34**	**.34**	**.34**
SVM-rbf$_{c}$	.35	**.55**	**.43**	**.21**	**.02**	**.04**	.00	.00	.00	.24	.01	**.03**	**.34**	**.03**	**.06**	**.39**	.65	.49	**.26**	**.21**	**.17**
SVM-rbf$_{d}$	**.35**	**.53**	**.42**	**.08**	.00	**.01**	**.33**	**.00**	**.01**	**.41**	**.03**	**.05**	.35	.11	**.17**	.38	.62	.47	**.32**	**.22**	**.19**
SVM-rbf$_{c+b}$	.35	**.60**	**.44**	.24	**.03**	**.05**	.00	.00	.00	.20	.01	.02	**.35**	**.04**	**.07**	**.41**	.64	.50	**.26**	**.22**	**.18**
SVM-rbf$_{c+d}$	.35	**.59**	**.44**	.20	.02	.04	.00	.00	.00	.24	.01	**.03**	**.32**	**.03**	**.06**	**.40**	.64	.49	**.25**	**.22**	**.18**
SVM-rbf$_{c+b+d}$	.35	**.62**	**.45**	.23	.02	.04	.00	.00	.00	.24	.01	**.03**	**.34**	**.04**	**.07**	**.43**	.64	**.51**	**.26**	**.22**	**.18**

Table A2. Precision (P), Recall (R) and F1 scores obtained by the baseline models using classic machine learning algorithms, with PCA. When a score is higher than its corresponding model in Table A1, it is highlighted in **bold**.