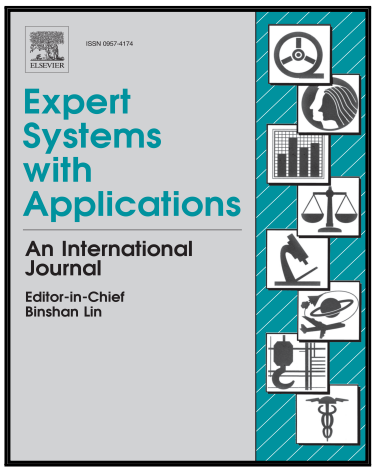


CDS: Collaborative Distant Supervision for Twitter Account Classification

Lishan Cui, Xiuzhen Zhang, A.K. Qin, Timos Sellis, Lifang Wu

PII: S0957-4174(17)30235-X
DOI: 10.1016/j.eswa.2017.03.075
Reference: ESWA 11233

To appear in: *Expert Systems With Applications*

Received date: 15 July 2016
Revised date: 30 March 2017
Accepted date: 31 March 2017

Please cite this article as: Lishan Cui, Xiuzhen Zhang, A.K. Qin, Timos Sellis, Lifang Wu, CDS: Collaborative Distant Supervision for Twitter Account Classification, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.03.075

Highlights

- Novel distant supervision-based approach to Twitter account classification
- Collaborative learning of distant supervision, active and semi-supervised learning
- Heuristics for automatically labelling Twitter accounts
- Generic strategy identifying false positives and false negatives from automatic labelling

CDS: Collaborative Distant Supervision for Twitter Account Classification

Lishan Cui[a], Xiuzhen Zhang[a,*], A. K. Qin[b], Timos Sellis[b], Lifang Wu[c]

[a] *RMIT University, Australia*
[b] *Swinburne University of Technology, Australia*
[c] *Beijing University of Technology, China*

Abstract

Individuals use Twitter for personal communication, whereas businesses, politicians and celebrities use Twitter for branding purposes. Distinguishing Personal from Branding Twitter accounts is important for Twitter analytics. Existing studies of Twitter account classification apply classical supervised learning, which requires intensive manual annotation for training. In this paper, we propose CDS (Collaborative Distant Supervision), a novel learning scheme for Twitter account classification that does not require intensive manual labelling. Twitter accounts are automatically labelled using heuristics for distant supervision learning. To achieve effective learning from heuristic labels, active learning is applied to identify and correct false positive labels, and semi-supervised learning is applied to further use false negatives missed by labelling heuristics for learning. Extensive experiments on Twitter data showed that CDS achieved high classification accuracy.

Keywords:
Twitter, classification, distant supervision, active learning, semi-supervised learning

*Corresponding author. Tel.: +61 3 9925 2774.
Email addresses: lishan.cui@rmit.edu.au (Lishan Cui), xiuzhen.zhang@rmit.edu.au (Xiuzhen Zhang), kqin@swin.edu.au (A. K. Qin), tsellis@swin.edu.au (Timos Sellis), lfwu@bjut.edu.cn (Lifang Wu)

1. Introduction

As much as Twitter is a social media platform for personal communication, Twitter is also a platform for branding exercises (Jansen et al., 2009). Ordinary individuals set up Twitter accounts to report personal activities and observations, whereas entities such as businesses, news agencies, celebrities, sportspersons and politicians use Twitter accounts to broadcast new products, services or news. According to a study by Brandwatch.com (2013), 63% of brands had multiple Twitter accounts in 2013, a sharp increase from 7% in 2011.

We categorise Twitter accounts into Personal and Branding accounts in a general sense. Such broad classification means that Branding accounts include spam accounts created by program robots. Although spam account detection is a separate problem (Grier et al., 2010; Laboreiro et al., 2011; Wang, 2010), our study is complementary to spam account detection. Filtering out Personal accounts as a first step can potentially improve spam detection accuracy. Moreover, the account profile features in our study can be used for spammer detection, complementary to the content-based features used in many existing spam detection approaches (Grier et al., 2010; Laboreiro et al., 2011; Wang, 2010).

Distinguishing Branding from Personal Twitter accounts is important for Twitter analytics. Profiling Twitter accounts in terms of gender (Bamman et al., 2012; Burger et al., 2011) and personality (Quercia et al., 2011) has attracted active research recently. An implicit but crucial assumption in these studies is that Twitter accounts under analysis are Personal accounts created by human individuals. If the Branding accounts were identified and removed in these studies, such data cleaning may have led these studies to different conclusions on gender and personality distribution. On the other hand, opinion mining and sentiment analysis on Twitter (Ghiassi et al., 2013; Kontopoulos et al., 2013; Pang and Lee, 2008; Zhou et al., 2014) are an important task for Twitter analytics. It is shown in a recent study (Yin et al., 2014) that, after the separation of Twitter data according to the account type, Branding accounts (businesses) show more prominent sentiment cycles in a day than Personal accounts (human

individuals). The result highlights the importance of distinguishing account types when using Twitter data to analyse mood patterns for humans (Golder and Macy, 2011).

Recent studies (De Choudhury et al., 2012; Oentaryo et al., 2015; Yan et al., 2013) typically formulated Twitter account classification as a standard supervised classification learning problem. Costly human labelling is required to acquire training data. Textual, social and temporal features are engineered from tweet textual content, user profile meta-data as well as link information for accounts and tweets.

Distant supervision (Craven et al., 1999; Mintz et al., 2009; Min et al., 2013; Surdeanu et al., 2012; Go et al., 2009; Zubiaga and Ji, 2013; Magdy et al., 2015a,b; Takamatsu et al., 2012) is a recent classification learning scheme not requiring manual labels – heuristics based on external knowledge sources are used to automatically label instances and to train classifiers (hence the name "distant supervision"). However, applying distant supervision to Twitter account classification faces several challenges. First, there is no existing external knowledge source that can be used to automatically label Twitter accounts. Secondly, the automatic labelling in distant supervision may generate false positives with wrong labels as well as negatives missed by the labelling heuristics. These issues are ignored in most distant supervision studies (Craven et al., 1999; Mintz et al., 2009; Go et al., 2009; Zubiaga and Ji, 2013; Magdy et al., 2015a,b). Only a few studies addressed these issues using application-specific external knowledge sources (Min et al., 2013; Surdeanu et al., 2012; Takamatsu et al., 2012), and the solutions can not be generalised to Twitter classification.

In this paper, we propose CDS (Collaborative Distant Supervision) for Twitter account classification without costly manual labelling. We specifically address the following research questions:

- Without any external knowledge source, are there heuristics for automatic labelling of Twitter accounts?
- How to identify true and false positives in automatic labels to improve

distant supervision?

- Are there false negatives (negatives missed by the labelling heuristics) that are still useful for training an effective classification model?

We propose heuristics based on the "following" relationship and retweeting behaviour of Twitter accounts to automatically label Branding and Personal accounts. To train an effective classification model with distant supervision, we further propose novel domain-independent generic strategies combining active learning and semi-supervised learning to identify false positives in automatic labels and false negatives missed by the labelling heuristics. To the best of our knowledge, our research is the first study that does not require intensive manual annotation for Twitter account classification. Our collaborative distant supervision approach also advances the current distant supervision learning.

We conducted experiments on Twitter data in April and May 2014 to benchmark CDS against state-of-the-art distant supervision-based classification algorithms. CDS significantly improved the accuracy of standard distant supervision. Especially when the heuristic labels were of low quality (43.36% of labels are correct), CDS improved the accuracy of distant supervision by a big margin (from 60.53% to 88.02%). When the heuristic labels were of reasonable quality (63.52% of labels are correct), CDS achieved an accuracy comparable with the standard classification model trained with complete manual labels (89.45% versus 90.90%). We applied CDS to perform exploratory analysis of a Twitter dataset of ∼0.7 million Twitter accounts and ∼44 million tweets for June 2014. The analysis confirmed the classification accuracy of CDS and revealed significant differences in Branding and Personal accounts for spreading information. Our analysis opens the door to further analysis of stakeholders on Twitter.

2. Related Work

Related work comes from three areas – Twitter account classification, distant supervision learning, and active and semi-supervised learning.

2.1. Twitter account classification

Existing studies on supervised learning-based Twitter account classification apply the standard supervised classification learning framework where costly human labels are used to train classification models (De Choudhury et al., 2012; Oentaryo et al., 2015; Yan et al., 2013). De Choudhury et al. (2012) studied the problem of classification of Twitter accounts into categories including organisations, journalists/media bloggers and ordinary individuals to determine who is using Twitter to broadcast information about current events. They proposed to build a standard supervised classification model based on human labels, using features including network/structural features, activity features, interaction features, named entities and topic distribution. Oentaryo et al. (2015) discriminated personal and organisation accounts using a comprehensive set of content, social, and temporal features in the classification model. Yan et al. (2013) classified Twitter accounts as closed (personal) accounts and open (business) accounts based on account profiles and follower distributions. The features of account profiles include clue keywords, telephone numbers, and detailed addresses.

Rather than building a discriminative classification model as in the previous three studies, Yin et al. (2014) proposed a probabilistic generative model based on Gaussian distribution of temporal, spatial and textual features to classify personal communication and public dissemination accounts.

2.2. Distant supervision classification learning

Distant supervision aims to avoid the costly manual labelling and automatically obtain labelled data for classification learning. Distant supervision was proposed in the biomedical domain (Craven et al., 1999) and recently gained popularity for text mining tasks (Mintz et al., 2009; Min et al., 2013; Surdeanu et al., 2012; Go et al., 2009; Zubiaga and Ji, 2013; Magdy et al., 2015a,b).

Distant supervision is applied to sentiment classification and topical classification on Twitter (Go et al., 2009; Zubiaga and Ji, 2013; Magdy et al., 2015a,b). Simple heuristics derived from either the tweet content or an external knowledge base are used for automatic heuristic labelling, but the issue of false positives

in heuristic labels is ignored. Go et al. (2009) classified tweets into positive and negative classes according to their sentiment polarity. For automatic labelling, they used emoticons from tweet contents as heuristics to label tweets as expressing positive or negative sentiment. Similarly, in some other studies (Zubiaga and Ji, 2013; Magdy et al., 2015a,b), tweets were classified according to their topical category. Zubiaga and Ji (2013) obtained topic labels for tweets following their URL links to the external knowledge source Open Directory Project. Magdy et al. (2015a,b) proposed to obtain heuristic labels for tweets following links to YouTube. In all these studies, heuristic labels are used in a standard supervised learning framework to train classification models.

Distant supervision is also applied to the relation extraction task, which aims to identify and extract semantic relations for entity pairs from natural language texts (Mintz et al., 2009; Min et al., 2013; Surdeanu et al., 2012). Mintz et al. (2009) described a standard distant supervision framework where the external knowledge source Freebase was used as supervision to label relations. Several studies (Min et al., 2013; Surdeanu et al., 2012) focused on the "false negative" issue in heuristic labels, where many real relations are missed by automatic labelling by the external knowledge base; a multi-instance classification learning model was proposed. Some other studies focused on addressing the "false positive" issue that some relations are assigned wrong labels by the knowledge base; a generative model was proposed by Takamatsu et al. (2012). The proposed approaches to identify false heuristic labels are based on specific external knowledge bases and are not generic solutions.

2.3. Active and semi-supervised learning

To achieve high learning accuracy with minimum human labelling cost, active learning strategies query the classification result of learning models to select the most difficult instances for human labelling. Commonly used active learning strategies include uncertainty sampling or boosted disagreement with query-by-committee (See Settles (2010) for an excellent survey). Semi-supervised learning aims to use limited labelled instances and a large number of unlabelled instances

for effective learning. Unlabelled instances are used for pseudo labelling (Bruzzone et al., 2006; Camps-Valls et al., 2007; Shi et al., 2011) or regularising the classification model (Nigam et al., 2000; Zhang et al., 2012; Sindhwani and Keerthi, 2006). Active learning and semi-supervised learning can also be combined for more effective classification (Munoz-Mari et al., 2012; Wan et al., 2015; Zhang et al., 2014).

3. Problem Statement

We distinguish two types of accounts on Twitter:

- *Branding* accounts are entities (natural or legal persons) that use Twitter for branding purposes. In other words, Branding accounts are associated with business, political or social goals. Branding accounts include companies (official corporate accounts), NGOs, charities, events, journalists/bloggers (freelance media professionals or news agencies), celebrities, politicians, sportsmen and spam robots.
- *Personal* accounts are natural persons who use Twitter for personal communication. In other words, Personal accounts are not associated with any specific goal. Personal accounts post tweets to express their opinions and describe updates on their professional life and daily life.

Our definitions of Branding and Personal accounts are aligned with the Open and Closed accounts by Yan et al. (2013), Public Dissemination and Personal Communication accounts by Yin et al. (2014) and Organisation and Personal accounts by Oentaryo et al. (2015). Note that there exist Twitter accounts used mainly for branding purposes but sometimes used for personal purposes. For example, the Twitter accounts of celebrities are used mainly for publicity but occasionally for disseminating information about their personal lives. In this case, these accounts are deemed Branding accounts.

Given Twitter accounts data, to classify the accounts into Branding and Personal account types, we formulate the task as a distant supervision classification problem. Specifically, given a large volume of Twitter account data, but

without any manual labels of Branding or Personal accounts, we aim to develop a classification model for account types. Consequently learning a model for Twitter account classification comprises the following steps:

1. Design heuristic rules based on domain knowledge to produce heuristic labels for accounts that will be used to train a classification model. In this sense, the domain knowledge encoded in heuristic rules provides distant supervision to learn the classification model. Such labelling rules are inherently heuristic and they may have low support – only accounts meeting the rules are labelled and many accounts are unlabelled. Moreover, the heuristic labelling rules can produce false labels.
2. Design strategies to address the issues of false heuristic labels and unlabelled instances in standard distant supervision.

Our heuristic labelling rules for Twitter accounts are described in Section 4, and the Collaborative Distant Supervision (CDS) learning algorithm is described in Section 5.

4. Heuristic rules for Automatic Labelling of Twitter Accounts

The heuristics for annotating instances are crucial for the classification accuracy of distant supervision. Generally heuristics that produce a large number of correct labels lead to an accurate classification model for distant supervision learning; the upper bound for heuristic labelling is that all instances under consideration are correctly labelled. As discussed earlier, textual conventions in tweets and external knowledge sources are used for Twitter sentiment classification (Go et al., 2009) and topical classification (Zubiaga and Ji, 2013; Magdy et al., 2015a,b). However, textual markers or external knowledge sources are not readily available for annotating Twitter accounts.

Past studies (De Choudhury et al., 2012; Oentaryo et al., 2015; Yan et al., 2013) show that in the Twitter sphere there are more Personal than Branding accounts, even though there has not been a general consensus on the ratio of Branding versus Personal accounts. We propose to apply this observation of an

uneven distribution of Branding and Personal accounts in the Twitter sphere to design simple threshold-based heuristic rules to label accounts. Based on this, we present the following observation:

Observation 1 *Given a population of Twitter accounts, there are more Personal accounts than Branding accounts.*

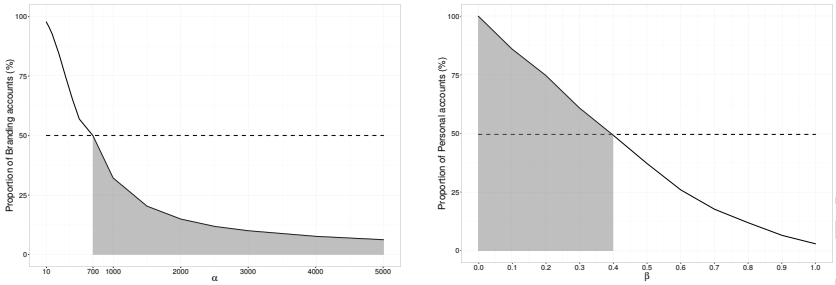
Based on Observation 1, given a population of Twitter accounts with account profile data and tweet posts, we have designed simple threshold-based heuristic rules to label accounts. The thresholds for heuristic labelling are set so that the total number of accounts labelled as Branding is at most 50% in the population, and that the Personal account is at least 50%.

It is shown in a previous study (Kwak et al., 2010) that Branding accounts, either celebrities (Ashton Kutcher, Britney Spears) or mass media (CNN breaking news, the New York Times, TIME), all have a high number of followers, irrespective of their account age. Indeed Branding accounts tend to gain popularity within a short period since the account's creation. Below, we present the heuristic for labelling Branding accounts.

Heuristic 1 Branding labelling heuristic *Let $f(u)$ denote the number of followers for account u. For a population of accounts, an account u is labelled Branding if $f(u) \geq \alpha$, where α is a threshold such that at most 50% of accounts in the population are labelled as Branding.*

The retweet mechanism plays a vital role for the information diffusion on Twitter (Kwak et al., 2010). It is observed by Kwak et al. (2010) that the top accounts ranked by retweets are Personal accounts. We generalise this observation to hypothesise that individual Personal accounts are more likely to retweet than Branding accounts. For labelling Personal accounts, we present the heuristic below.

Heuristic 2 Personal labelling heuristic *Let $rt(u)$ denote the retweet rate for account u, which is the portion of retweets in the tweets of u. For a population*



(a) $f(u) \geq \alpha$ (b) $rt(u) \geq \beta$

Figure 1: Account label proportion at different thresholds for Branding heuristic $f(u) \geq \alpha$ and Personal heuristic $rt(u) \geq \beta$ for the population of 691,391 accounts in March 2014.

of accounts, an account u is labelled Personal if $rt(u) \geq \beta$*, where* β *is a threshold such that at least 50% of accounts in the population are labelled as Personal.*

We collected a population of Twitter accounts in March 2014, which consists of 691,391 Twitter account profiles and 12.9 million tweets. Figure 1 depicts the proportion of labelled accounts in the population at various threshold settings for labelling. We can see from Figure 1(a) that when $\alpha \geq 700$, the rule "$f(u) \geq \alpha$" labels $< 50\%$ of accounts in the population as Branding. On the other hand, Figure 1(b) shows that for $\beta \leq 0.4$, the rule "$rt(u) \geq \beta$" labels $\geq 50\%$ of accounts in the population as Personal. Consequently $\alpha \geq 700$ and $\beta \leq 0.4$ can be used to set the thresholds for the heuristic labelling rules for Branding and Personal accounts respectively. It should be noted that generally another population of sufficient size generates similar threshold values for labelling.

The Branding and Personal labelling rules are heuristics in nature and can produce false labels. We thus call the true and false labels generated by the labelling rules *true positives* and *false positives* respectively. On the other hand, the accounts that do not meet the labelling rule thresholds are not labelled and are *negatives*.

We performed post-hoc analysis with a random sample from the population of accounts in March 2014 to evaluate the level of true and false positives, and false negatives in the heuristic labels by the labelling rules. We manually labelled

Table 1: Number of labels (for a sample of 2000 accounts) and percentage of true positive labels for α and β settings

α and β	# of heuristic labels	# of true positives (%)
$\alpha = 700$, $\beta = 0.1$	724	40.16%
$\alpha = 1000$, $\beta = 0.3$	749	36.81%
$\alpha = 1500$, $\beta = 0.3$	763	39.68%
$\alpha = 2000$, $\beta = 0.3$	764	40.58%
$\alpha = 2000$, $\beta = 0.4$	656	32.69%
$\alpha = 2500$, $\beta = 0.2$	878	49.21%
$\alpha = 2500$, $\beta = 0.3$	769	41.47%
$\alpha = 3000$, $\beta = 0.4$	638	33.17%

a random sample of 2000 accounts where 421 accounts are labelled Branding and 1038 accounts are labelled Personal, and 541 accounts are either suspended by Twitter or do not have public profiles. We performed the labelling rules $f(u) \geq \alpha$ and $rt(u) \geq \beta$ to label Branding and Personal accounts respectively. When an account meets both rules it is deemed unlabelled. When α ranges in 700..3000 with a step of 500, and β ranges in 0.1..0.4 with a step of 0.1, there are 24 settings for (α, β). Different settings of α and β produce different number of heuristic labels, as well as different levels of true and false positives and false negatives, as shown in Table 1.

The issues of false positives and false negatives in heuristic labels need to be addressed for effective distant supervision learning. Strategies adopted by the CDS algorithm for addressing these issues are described in the next section. Moreover our experiments show that CDS is not very sensitive to the labelling rule threshold settings and can achieve effective Twitter account classification.

5. Collaborative Distant Supervision for Classifying Twitter Accounts

We propose CDS, a collaborative distant supervision classification model that applies active learning and semi-supervised learning to identify false posi-

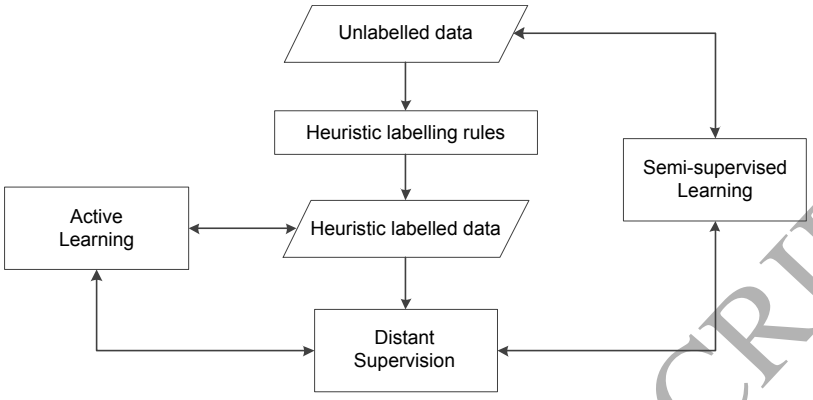


Figure 2: The collaborative learning framework

tives and false negatives in heuristic labelling. The collaborative learning framework is as shown in Figure 2. Within the learning framework, the unlabelled data are first heuristically labelled to acquire heuristic labels for distant supervision learning. Distant supervision learning then interacts with active learning and semi-supervised learning to identify false positives and false negatives from heuristic labelling and to improve learning efficiency. Specifically an active learning query strategy is applied to actively identify instances with false heuristic labels and rectify their labels. In other words, instances that have inconsistent heuristic and classification labels and that are most unreliably classified by the classification model are false positives in heuristic labels, and these instances are manually labelled to rectify their labels. On the other hand, semi-supervised learning is applied to enrich the distant supervision classification model. Specifically, unlabelled instances consistently classified by the model based on heuristic labels and the model based on manual labels are deemed false negatives by heuristic labelling and are added to enlarge the training population. As a result, the final collaborative distant supervision model is based on limited manual labels by active learning and the unlabelled instances with acquired labels by semi-supervised learning from instances with consistent heuristic and classification labels.

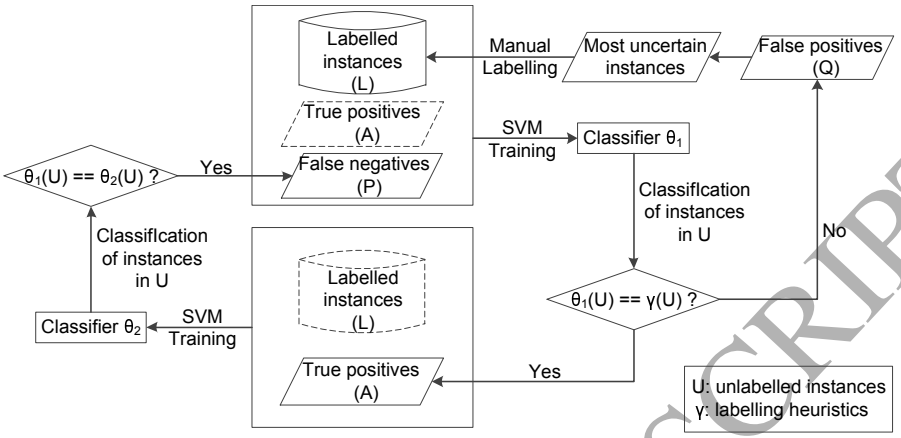


Figure 3: The flowchart for CDS (dashed line shapes are replicates for clarity)

5.1. The CDS algorithm

The flowchart for CDS is shown in Figure 3. CDS is an iterative learning process. Input to CDS includes a set of unlabelled instances U and the labelling heuristics γ (Section 4). The base classification model for CDS is SVM (Chang and Lin, 2011), which is commonly used for document classification and active learning. Applying γ to instances in U generates two sets: positives labelled by γ and negatives otherwise. Instances with heuristic labels consistent with the classification results by classifier θ_1 (explained next) are true positives (denoted as A), and are false positives otherwise (denoted as Q). Applying active learning, a subset of false positives $Q' \subset Q$, namely those most uncertainly classified by θ_1, are manually labelled (denoted as L) to correct the false heuristic labels. Classifier θ_2 is trained from $L \cup A$. On the other hand, negatives that are not heuristically labelled but are consistently classified by θ_1 and θ_2 are false negatives (denoted as P) – in the sense that they are falsely unlabelled. Classifier θ_1 is trained from $L \cup A \cup P$ and is the final output classifier when the iterative learning process finishes.

The pseudocode for CDS is shown in Algorithm 1. Given a set U of given unlabelled instances and labelling heuristics γ, the learning process is bootstrapped by applying γ to label instances in U (Line 1). All heuristic labelled

Algorithm 1 The CDS Algorithm

Input: U: input instances; γ: labelling heuristics; q: #instances for active learning.

Output: Classifier θ_1.

1: $P \leftarrow \{d \in U | \gamma(d)\}$ *;; initialise P*
2: $L \leftarrow \phi$ *;; L: instances with human labels*
3: **repeat**
4: $\theta_1 \leftarrow$ SVM-train $(L \cup A \cup P)$ *;; θ_1: final classifier to output*
5: $A \leftarrow \{d \in U | \theta_1(d) == \gamma(d)\}$ *;; A: true positive heuristic labels*
6: $Q \leftarrow \{d \in U | \theta_1(d) != \gamma(d)\}$ *;; Q: candidate false positive labels*
7: Select q instances $Q' \subset Q$ with highest uncertainty *;; active learning*
8: Manually label instances in Q' *;; Q': false positive heuristic labels*
9: $L \leftarrow L \cup Q'$
10: $U \leftarrow U - L - A$
11: $\theta_2 \leftarrow$ SVM-train $(L \cup A)$ *;; θ_2 from true positives and the manual labelled*
12: $P \leftarrow \phi$ *;; P: classification labels for training θ_1*
13: **for** each instance $d \in U$ **do**
14: **if** $\theta_1(d) == \theta_2(d)$ **then**
15: $P \leftarrow P \cup \{d\}$
16: **end if**
17: **end for** *;; Lines 13–17 identify false negatives*
18: **until** P=U
19: Return the final classification model θ_1

instances become the initial set P, and the set of instances with human labels $L = \phi$ (Line 2). With the iterative learning process, for each iteration (Line 3 to Line 18), two classifiers θ_1 and θ_2 are trained. Classifier θ_1 is the final classifier trained on $L \cup A \cup P$. Classifier θ_2 (Line 11) is trained on $L \cup A$. (Line 5). Over iterations, the unlabelled instance set U becomes smaller. The iteration finishes when all remaining instances in U are identified as false negatives consistently classified by θ_1 and θ_2, or $P = U$.

In the active learning step, to select candidate query instances Q based on the θ_1 classification results, heuristic labels are used to assist selecting candi-

Table 2: Features for Twitter account classification

Account profile features	
ScreenNameLength	length of screen_name
UserNameTokens	number of tokens of user_name
NameCommon	$\frac{\#common\ chars\ in\ screen\_name\ and\ user\_name}{ScreenNameLength}$
FollowingCount	number of followings
FollowRatio	$\frac{\#followers}{\#followings}$
FollowerMean	mean of #followers over time
FollowerStdDev	deviation of #followers over time
FollowingMean	mean of #followings over time
FollowingStdDev	deviation of #followings over time
ListCount	# public lists as a member
ListMean	mean of #lists over time
ListStdDev	deviation of #lists over time
FavouriteCount	#tweets favourited
FavouriteRate	$\frac{FavouriteCount}{AccountAge}$
TweetCount	total number of tweets
TweetCountRate	$\frac{total\ number\ of\ tweets}{AccountAge}$
DescriptionTokens	#tokens in description
Tweet content features	
TweetCount	#tweets
RetweetUserMention	proportion of retweets of unique user mentions
UserMention	proportion of tweets with user mentions
UserMentionUnique	proportion of unique user mentions
URLCount	proportion of URLs
URLUnique	proportion of unique URLs
HashtagCount	proportion of hashtags
HashtagUnique	proportion of unique hashtags
EmoticonCount	proportion of emoticons
Term	proportion of tweets with self-reference terms
SourceUnique	#unique sources

date query instances for human labelling and train the classifier θ_2. Note that our active learning process assisted by heuristic labels is very different from conventional active learning, where the most uncertain instances are selected according to the classification result (Settles, 2010). Our candidate instances are those instances with unreliable heuristic labels, or false positive heuristic labels different from the classification labels. These candidate instances are further sampled for a pre-set of q (by default 10) instances for human labelling (Line 7). To select q instances, margin sampling that selects instances close to the SVM separating hyperplane is applied in combination with the kernel k-means clustering algorithm to increase diversity (Wan et al., 2015).

In the semi-supervised learning step, "good" unlabelled instances (false negatives) are identified and used to enrich the training population (Lines 13 to 17). Unlabelled instances in $L \cup A$ are with human labels (L) or reliable true positive heuristic labels (A) and are used to train θ_2. All instances in U without human labels are classified by θ_2. Only instances that are consistently labelled by θ_1 and θ_2 are included in P to train θ_1, the final classifier to output. As a result the final classifier θ_1 is trained from false positives selected by active learning and with rectified manual labels L, true positives A, and false negatives P.

With active learning in CDS, the small number of human labels can guide the learning process to derive an accurate classifier, despite an initial set of heuristic labels that includes false positives. On the other hand, semi-supervised learning can enrich distant supervision using unlabelled instances with consistent acquired classification labels. As will be shown in our experiments, manual labelling only 10 instances at each active learning step can achieve very effective classification, helps the learning process converge.

5.2. The features

Previous studies show that Twitter account profile features and tweet content features are likely linked to account types (Oentaryo et al., 2015; Yan et al., 2013), including content, social and temporal features as well as account profile and follower distribution features. The features used in the CDS algorithm

are listed in Table 2, including 17 account public profile features and 11 tweet content features. The tweet content features need to be computed from the tweets for the Twitter accounts. Semantics of most features are self-explanatory. We further explain a few features as follows:

- *ListCount*: The number of public lists of which an account is a member. A Twitter account can create his/her own lists or subscribe to lists created by others. Viewing a list timeline will show a stream of tweets from the users on that list. The list for a Twitter account is like his/her social circle.

- *EmoticonCount*: The proportion of tweets containing emoticons. Emoticons are representation of facial expression to express tweet author's emotions. We include commonly used emoticons :), :(, :D, :-(and :-).

- *Term*: The proportion of tweets containing self-reference terms. We consider self-reference terms "I", "me" and "my". Individuals are expected to often use these self-reference terms in their tweets.

Note that many profile features are based on the "Following" or "Follower" links of accounts. It is shown that these social link features are linked to account types. For example, the feature "FollowerFollowingRatio" is about the reciprocity of social links for accounts. As reported by Kwak et al. (2010), Twitter accounts demonstrate different levels of reciprocity – celebrity accounts have many followers but much less followings.

The language characteristics of tweets are also indicative of account behaviour and linked to account types. Generally speaking Personal accounts tend to use more emoticons in their tweets, as they are expected to use Twitter to express opinions or sentiment. In contrast, Branding accounts post tweets for branding exercises like marketing or customer service, and are expected to use emoticons far less frequently. Moreover, as individuals mainly post tweets for personal use, naturally they use self-reference terms like "I", "me" or "my" more often than Branding accounts.

6. Experiments

We implemented the CDS algorithm as described in Algorithm 1 with LibSVM (Chang and Lin, 2011) as the base learner. Settings for LibSVM are as follows: the default C-SVC type of SVM, kernel type is radial basis function, settings of cost as 2^{14} and gamma as 2^{-6}. Note that these parameter settings are not to optimise the performance of SVM, as parameter tuning is not the focus of our study.

CDS is benchmarked against four state-of-the-art distant supervision-based learning schemes:

- DS (Go et al., 2009) — A standard distant supervision algorithm based on SVM. DS was widely used for Twitter classification tasks such as sentiment classification and account classification.
- DS^+ (Min et al., 2013)— An enhanced distant supervision algorithm based on SVM where semi-supervised learning is applied to further use the negatives missed by heuristic labelling to enhance distant supervision. DS^+ was shown to outperform DS for the task of relation extraction from natural language texts.
- GBMDS (Oentaryo et al., 2015) — A gradient boosting machine-based distant supervision algorithm using a comprehensive set of content, social and temporal features. GBMDS is a representative recent study on Twitter account classification.
- DeepDS (Dhingra et al., 2016) — A distant supervision algorithm where deep learning is applied to learn character-based distributed vector representations from tweets as features for learning. The character-based distributed vector representation can overcome the challenge of out-of-vocabulary words and form high level representation for character sequences. It was shown recently that such representation can improve the accuracy of Twitter classification such as sentiment classification and topic classification.

Table 3: The dataset statistics

	#Tweets	#Accounts	#Active accounts
April 2014	43,942,338	12,799,936	704,864
May 2014	44,279,374	12,695,887	706,058
June 2014	44,378,334	12,540,919	732,778
Total unique	132,600,046	23,173,552	1,327,835

6.1. The datasets

We crawled tweets in April, May and June 2014 using the Twitter Stream API. We used datasets of April and May for development and evaluation of our classification model, and used the June dataset to apply the trained model to perform exploratory analysis. We compiled over 132.6 million tweets by 23.2 million accounts where the stated language for their corresponding Twitter account profiles is English. Statistics of the datasets are shown in Table 3. As there are accounts that posted tweets in all months, the total number of unique accounts in three months are less than the number of accounts in each month separately. There are in total over 1.3 million active accounts, which are live Twitter accounts not suspended by Twitter and that posted at least 10 tweets in each month. The threshold of 10 tweets is to ensure that tweet content profile features can be computed. Later discussions will be based on active accounts.

We need a Twitter data with account type annotations as ground truth to evaluate the performance of different algorithms. To the best of our knowledge there does not exist such a public Twitter dataset. We therefore constructed a ground truth dataset as follows: We randomly selected 2,500 accounts based on the crawled tweets for April and May 2014. The 129 accounts suspended by Twitter or not having public profiles were discarded. Three volunteers were recruited to annotate the rest 2317 accounts as Branding or Personal accounts. Annotators were given the definitions for Branding and Personal accounts in Section 3. They were instructed to check the account public profiles to label accounts, and were shown the public profiles of some known Branding and Per-

Table 4: The Accuracy of CDS versus the other approaches

Threshold	TP (%)	CDS	DS	DS$^+$	GBMDS	DeepDS
$\alpha 1000 \beta 0.3$	43.36%	0.8802	0.6053	0.6636	0.5887	0.6669
$\alpha 1500 \beta 0.3$	50.11%	0.8879	0.6182	0.7177	0.5887	0.7355
$\alpha 2000 \beta 0.3$	51.75%	0.8925	0.6245	0.7373	0.5897	0.7623
$\alpha 2500 \beta 0.3$	52.76%	0.8937	0.6314	0.7504	0.5984	0.7905
$\alpha 2500 \beta 0.2$	63.52%	0.8945	0.7229	0.7888	0.6939	0.8199

sonal Twitter accounts. The pairwise Kappa statistics were employed to assess the level of annotation agreement among the three annotators. The resulting Kappa statistics were in the range $0.43 \sim 0.58$, indicating acceptable moderate level of agreement (Landis and Koch, 1977). Majority vote was used to decide the final label for instances. In the end 2371 accounts included 425 Branding and 1946 Personal accounts.

6.2. Performance of CDS

Experiments were run on the 2371 Twitter accounts with ground truth labels to evaluate the performance of CDS against DS, DS$^+$, GBMDS, and DeepDS. We ran each algorithm 20 times where each run used randomly selected 80% of data for training (without using their labels) and the rest 20% for testing. We evaluated the accuracy of CDS with respect to the quality of heuristic labels, in comparison to DS, DS$^+$, GBMDS, and DeepDS. Our hypothesis is that the performance of DS, DS$^+$, GBMDS, and DeepDS very much depends on the quality of heuristic labels. In contrast, CDS with limited human labels by active learning and semi-supervised learning can mitigate the false positives and false negatives in heuristic labels and produce high quality classification.

Table 4 lists the classification accuracy of CDS in contrast to the other four approaches at different threshold settings for heuristic labelling. Recall that from Table 1, different threshold settings resulted in different levels of true positive heuristic labels ranging from 43.36% to 63.52%. Our experiment

Table 5: The Precision, Recall and F_1 of CDS versus the other approaches

Threshold	Alg	Branding			Personal		
		Prec	Recall	F_1	Prec	Recall	F_1
$\alpha 1000 \beta 0.3$	CDS	0.6631	0.6847	0.6737	0.9305	0.9229	0.9267
	DS	0.2571	0.6365	0.3663	0.8831	0.5985	0.7134
	DS^+	0.2638	0.4888	0.3427	0.8629	0.7018	0.7740
	GBMDS	0.2376	0.5853	0.3380	0.8668	0.5895	0.7017
	DeepDS	0.2587	0.4606	0.3313	0.8582	0.7120	0.7783
$\alpha 1500 \beta 0.3$	CDS	0.7078	0.6418	0.6732	0.9233	0.9416	0.9324
	DS	0.2641	0.6312	0.3723	0.8844	0.6154	0.7258
	DS^+	0.3130	0.4794	0.3787	0.8714	0.7698	0.8175
	GBMDS	0.2376	0.5853	0.3380	0.8668	0.5895	0.7017
	DeepDS	0.2885	0.3247	0.3056	0.8484	0.8253	0.8367
$\alpha 2000 \beta 0.3$	CDS	0.7391	0.6218	0.6754	0.9202	0.9517	0.9356
	DS	0.2651	0.6171	0.3708	0.8823	0.6261	0.7324
	DS^+	0.3370	0.4771	0.3950	0.8743	0.7942	0.8323
	GBMDS	0.2370	0.5800	0.3365	0.8658	0.5918	0.7030
	DeepDS	0.3053	0.2559	0.2784	0.8430	0.8730	0.8578
$\alpha 2500 \beta 0.3$	CDS	0.7610	0.5947	0.6677	0.9155	0.9590	0.9368
	DS	0.2700	0.6188	0.3759	0.8841	0.6342	0.7386
	DS^+	0.3534	0.4671	0.4024	0.8747	0.8123	0.8423
	GBMDS	0.2403	0.5735	0.3387	0.8664	0.6039	0.7117
	DeepDS	0.3640	0.2271	0.2797	0.8440	0.9136	0.8774
$\alpha 2500 \beta 0.2$	CDS	0.7835	0.5694	0.6595	0.9113	0.9656	0.9376
	DS	0.3247	0.5035	0.3948	0.8767	0.7708	0.8204
	DS^+	0.4069	0.3900	0.3983	0.8680	0.8760	0.8720
	GBMDS	0.2887	0.4829	0.3614	0.8677	0.7400	0.7987
	DeepDS	0.4914	0.1482	0.2278	0.8386	0.9667	0.8981

examined the performance of CDS against the other four algorithms with respect to heuristic labels of varying quality. CDS achieved stable accuracy hovering at 0.89 (0.8802∼0.8945) when the number of true positive heuristic labels increased from 43.36% to 63.52%. On the contrary, the other four approaches showed accuracy very much dependent on the quality of heuristic labels. The threshold setting of $\alpha 1000 \beta 0.3$ produced low quality heuristic labels with 43.36% true positives. As a result DS and GBMDS had very low accuracies, and DS^+ and DeepDS got better accuracies of 0.6636 and 0.6669 respectively. The threshold setting of $\alpha 2500 \beta 0.2$ produced high quality heuristic labels with a true positive heuristic label rate of 63.52%, and so all four algorithms got better classification

Table 6: Performance of CDS in terms of number of instances to label in each iteration

#inst (q)	#iterations	Accuracy
5	60.50	89.16%
7	49.00	89.39%
10	37.05	89.37%
15	30.55	89.89%
20	25.30	90.03%
30	20.00	90.22%
50	15.00	90.42%
70	12.40	90.54%
90	10.50	90.75%
100	9.60	90.56%

accuracies. Following the recommendation by Salzberg (1997), McNemar test was applied to compare CDS against the other four algorithms, and the CDS accuracy was significantly better (p-value $<< 0.001$).

Considering the imbalanced distribution for Branding and Personal classes, we further evaluated the class-specific performance of CDS versus DS, DS$^+$, GBMDS and DeepDS, and the results are shown in Table 5. Performance metrics include Precision, Recall, and the F_1 measure for the Branding and Personal classes respectively. In our experiment, CDS showed consistently better performance than the other approaches across all evaluation metrics for both Branding and Personal classes (CDS and DeepDS had comparable Recall scores for the Personal class at the setting of $\alpha 2500 \beta 0.2$). For the minority Branding class, while CDS achieved F_1 scores in the range of 0.6595..0.6754, the other four approaches had much lower F_1 scores. For the Personal class, while CDS achieved F_1 scores in the range of 0.9267..0.9376, the other approaches obtained modest F_1 scores. It can be seen that CDS can achieve accurate classification for each class despite the imbalanced class distribution. In contrast, the other four algorithms performed poorly on the minority Branding class. The robust

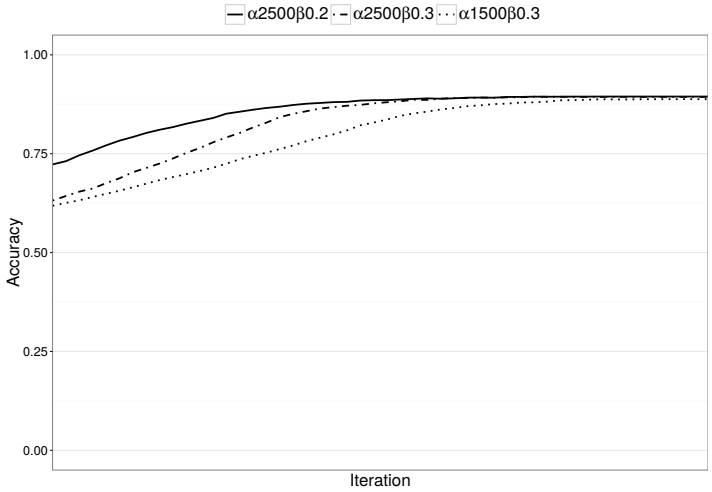


Figure 4: The learning curve for CDS with different Labelling rules

performance of CDS in the presence of class imbalance may be explained by its hybrid active and semi-supervised learning strategy. In particular active learning selects "informative" false positives from heuristic labelling and semi-supervised learning selects false negatives missed by heuristic labelling that are most likely minority-class instances. As a result the hybrid strategy can mitigate class imbalance and achieve effective classification (Zhang et al., 2017).

6.3. Sensitivity analysis of CDS

The only parameter of CDS is q, the number of instances to manual label for active learning at each iteration. We measured the rate of convergence and accuracy of CDS for different settings of q in the range of 5 to 100 with the heuristic labelling rule settings of ($\alpha = 2500$, $\beta = 0.3$), and the results are as shown in Table 6. The first column of the table shows the number of iterations (averaged over ten folds) for CDS to converge at different settings of q. As expected, with a larger q value, CDS converges more quickly. For small values of q, a small increase in q results dramatic reduction in number of iterations. For example when q increases from five to seven, the average number of iterations for the algorithm to converge reduces sharply from 60.50 to 49.00. This suggests

that our strategy of using heuristic labels to guide the selection of instances for active learning has been very effective. The second column of the table shows that CDS has stable classification performance across different settings of q – when q increases from 5 to 100, the classification accuracy is between 89.16% and 90.75%. These results demonstrate that a small number of manually labelled instances lead to stable performance for CDS.

We further analyse the learning curve for CDS – how quickly it converges. Figure 4 plots the accuracy of CDS in terms of the number of iterations when $q = 10$ and with the heuristic labelling rule of ($\alpha = 2500$, $\beta = 0.2$), ($\alpha = 2500$, $\beta = 0.3$), and ($\alpha = 1500$, $\beta = 0.3$). It can be seen that with "$f(u) \geq 2500$" for Branding and "$rt(u) \geq 0.2$ " for Personal ($\alpha = 2500$, $\beta = 0.2$), CDS becomes stable at iteration 30, reaching an accuracy of 89.45%. It is important to note that CDS achieves this performance with 30 iterations, which means that only 300 instances are labelled. Even with the heuristic labelling rule of ($\alpha = 1500$, $\beta = 0.3$), CDS becomes stable at iteration 38, which means that only 380 instances are labelled. This shows that our strategy of using heuristic labels to guide the active learning process has been especially effective.

7. Exploratory Analysis for Account Types

Twitter is a powerful platform for information dissemination (Kwak et al., 2010). Our automatic classification of Twitter account types allows deeper characterisation of differences in Branding and Personal accounts for disseminating information on Twitter. We applied the CDS classifier trained on the 2371 random accounts in April and May 2014 to classify ∼0.73 million Twitter accounts in June 2014, and then analysed their 44 million tweets posted in June 2014 (details in Table 3) to characterise differences of Branding and Personal accounts in terms of topic distribution and participation of topics. Our goal is to use a sizeable dataset completely separate from the experimental dataset for exploratory analysis.

For the account classification result, we found that among the ∼0.73 million

(a) Branding accounts (b) Personal accounts

Figure 5: The hashtag frequency cloud

accounts in June 2014, about 332,000 (45.4%) accounts are Branding accounts and about 400,000 (54.6%) accounts are Personal accounts. For validation we manually labelled 1000 random accounts as before (Section 6.1). Comparison of the manual labels and automatic labels by CDS shows that CDS yields high accuracy of ∼86%, demonstrating that CDS is a robust classifier for large-scale exploratory analysis. We next present results for analysing the 44 million tweets extracted in June 2014 for these accounts.

To understand the differences between Branding and Personal accounts in disseminating information on Twitter, we investigated the tweet behaviour and topic characteristics for the two account types. Around 7.9 million tweets were posted in total by 332,648 Branding accounts and around 6.8 million tweets were posted in total by 400,130 Personal accounts. Branding accounts, although a smaller proportion of the Twitter sphere, posted more tweets. Regarding using the retweet mechanism for information diffusion, 34.98% of tweets by Branding accounts are retweets whereas 41.83% tweets by Personal accounts are retweets. So in general Personal accounts tend to use the retweet mechanism more than Branding accounts.

Hashtags signify topics of information spread on Twitter (Romero et al., 2011). We analysed the popular hashtags in tweets by Branding and Personal accounts with a frequency of at least 100 in June 2014. Figure 5 plots the

hashtag cloud for the popular hashtags of Branding and Personal accounts after removing spam hashtags such as #porn, #sex, which have an absurdly high frequency. The hashtag cloud in Figure 5(a) is for Branding accounts and that in Figure 5(b) is for Personal accounts, where more frequent hashtags appear larger and bolder. As each cloud is relative to different hashtag population, the size and boldness of hashtags are not comparable across the two clouds. The frequency distribution for hashtags follows the Zipf's law, the power of Branding accounts is 0.94, and the power of Personal accounts is 0.83.

Comparing Figure 5(a) and Figure 5(b) it can be seen that hashtags are more evenly distributed for Personal accounts than Branding accounts. For Personal accounts there is a significant number of hashtags with similar frequencies. In contrast for Branding accounts there is a small number of hashtags with distinctly high frequency whereas most hashtags have low frequencies. We further analysed the top hashtags for each account type and annotated these hashtags by searching Twitter and other Web sources. From Figure 5 (a), the top three hashtags for the Branding accounts are

#gameinsight — Game Insight, a mobile game developer;
#android – Android, a mobile operating system by Google;
#androidgames — Android games.

Interestingly all these three hashtags are about product brands. From Figure 5(b), the top three hashtags for the Personal accounts are

#mileyformmva — Miley (Cyrus) for MMVA (Much Music Video Awards);
#android – Android;
#selenaformmva — Selena (Gomez) for MMVA.

The first two popular hashtags for Personal accounts are about celebrities (Miley Cyrus and Selena Gomez are popular singers). Our examination of the related tweet contents reveals that the tweet posts are individuals expressing their sentiment towards the two celebrities. This result of the topics aligns perfectly with

our definition for the Branding and Personal accounts. Note that the hashtag #android are popular for both Branding and Personal accounts.

We further analysed how different account types use hashtags. The Branding account uses a hashtag 3.49 times in the month on average whereas a Personal account uses a hashtag 1.82 times in the month on average. This result is in line with the theory of "persistence" of hashtags – repeated usage of the same hashtag helps spreading information (Romero et al., 2011). Moreover our analysis shows that Branding accounts are more persistent at spreading information for branding purposes.

8. Conclusions

Classifying Twitter accounts into Branding or Personal is important for uncovering the stakeholders on Twitter. Existing studies based on supervised learning require costly manual labelling. In this paper, motivated to reduce human labelling, we proposed CDS based on distant supervision. To reduce the false positives and false negatives by heuristic labelling, CDS adopts a collaborative learning scheme combining distant supervision with active and semi-supervised learning. Experiments on Twitter data demonstrated that CDS significantly improved the classification accuracy of distant supervision, especially when the heuristic labels are of low quality. We applied CDS to analyse ∼0.7 million Twitter accounts and 44 million tweets and found significant differences for Branding and Personal accounts on spreading information on Twitter.

For future work we will study developing our collaborative learning framework into online training and classification for the Twitter data stream. We will also explore a more general collaborative learning framework for other classification problems.

Acknowledgements

The authors thank the Victorian government, Australia and Lexer Pty Ltd, especially Mr Aaron Wallis, for supporting this research. The authors thank

Mr. Cheng Yu for the early work leading to this research and Dr. David Savage for checking the English of the paper.

References

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2012). Gender in Twitter: Styles, stances, and social networks. *Computing Research Repository*.

Brandwatch.com (2013). How Many Twitter Accounts Should a Brand Have? Retrieved from http://www.brandwatch.com/2013/08/the-rise-of-the-multiple-twitter-accounts/. Accessed: 19-July-2015.

Bruzzone, L., Chi, M., and Marconcini, M. (2006). A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363–3373.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Camps-Valls, G., Marsheva, T. V. B., and Zhou, D. (2007). Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3044–3054.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27.

Craven, M., Kumlien, J., et al. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86.

De Choudhury, M., Diakopoulos, N., and Naaman, M. (2012). Unfolding the event landscape on Twitter: classification and exploration of user categories.

In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244.

Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., and Cohen, W. W. (2016). Tweet2vec: Character-based distributed representations for social media. In *Proceedings of ACL*, pages 269–274.

Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Golder, S. A. and Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Grier, C., Thomas, K., Paxson, V., and Zhang, M. (2010). @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and Communications Security*, pages 27–37.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of Twitter posts. *Expert systems with applications*, 40(10):4065–4074.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Laboreiro, G., Sarmento, L., and Oliveira, E. (2011). Identifying automatic posting systems in microblogs. In *Proceedings of Portuguese Conference on Artificial Intelligence*, pages 634–648.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Magdy, W., Sajjad, H., El-Ganainy, T., and Sebastiani, F. (2015a). Bridging social media via distant supervision. *Social Network Analysis and Mining*, 5(1):1–12.

Magdy, W., Sajjad, H., El-Ganainy, T., and Sebastiani, F. (2015b). Distant Supervision for Tweet Classification Using YouTube Labels. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 638–641.

Min, B., Grishman, R., Wan, L., Wang, C., and Gondek, D. (2013). Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the HLT-NAACL*, pages 777–782.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Association for Computational Linguistics(ACL)*, pages 1003–1011.

Munoz-Mari, J., Tuia, D., and Camps-Valls, G. (2012). Semisupervised classification of remote sensing images with active queries. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):3751–3763.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2-3):103–134.

Oentaryo, R. J., Low, J.-W., and Lim, E.-P. (2015). Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts. In *Proceedings of the Advances in Information Retrieval*, pages 465–476.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our Twitter profiles, our selves: Predicting personality with Twitter. In *Proceedings of the International Conference on SocialCom and PASSAT*, pages 180–185.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Shi, L., Ma, X., Xi, L., Duan, Q., and Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications*, 38(5):6300–6306.

Sindhwani, V. and Keerthi, S. S. (2006). Large scale semi-supervised linear svms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 477–484.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the Association for Computational Linguistics(ACL)*, pages 455–465.

Takamatsu, S., Sato, I., and Nakagawa, H. (2012). Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the Association for Computational Linguistics(ACL)*, pages 721–729.

Wan, L., Tang, K., Li, M., Zhong, Y., and Qin, A. (2015). Collaborative active and semisupervised learning for hyperspectral remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2384–2396.

Wang, A. H. (2010). Detecting spam bots in online social networking sites: a machine learning approach. In *Proceedings of the 24th Annual IFIP Conference on Data and Applications Security and Privacy*, pages 335–342.

Yan, L., Ma, Q., and Yoshikawa, M. (2013). Classifying Twitter users based on user profile and followers distribution. In *Proceedings of International Conference on Database and Expert Systems Applications*, pages 396–403.

Yin, P., Ram, N., Lee, W.-C., Tucker, C., Khandelwal, S., and Salathé, M. (2014). Two sides of a coin: Separating personal communication and public dissemination accounts in Twitter. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 163–175.

Zhang, X., Li, Y., Kotagiri, R., Wu, L., Tari, Z., and Cheriet, M. (2017). KRNN: k rare-class nearest neighbour classification. *Pattern Recognition*, 62:33–44.

Zhang, X., Zhou, Y., Bailey, J., and Ramamohanarao, K. (2012). Sentiment analysis by augmenting expectation maximisation with lexical knowledge. In *Proceedings of the Web Information Systems Engineering*, pages 30–43.

Zhang, Y., Wen, J., Wang, X., and Jiang, Z. (2014). Semi-supervised learning combining co-training with active learning. *Expert Systems with Applications*, 41(5):2372–2378.

Zhou, Z., Zhang, X., and Sanderson, M. (2014). Sentiment analysis on Twitter through topic-based lexicon expansion. In *Proceedings of the Australasian Database Conference*, pages 98–109.

Zubiaga, A. and Ji, H. (2013). Harnessing web page directories for large-scale classification of Tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 225–226.