# Real-time Detection and Sorting of News on Microblogging Platforms

**Wenting Tu** **David W. Cheung** **Nikos Mamoulis** **Min Yang** **Ziyu Lu**
Department of Computer Science
The University of Hong Kong
Pokfulam, Hong Kong
{wttu,dcheung,nikos,myang,zylu}@cs.hku.hk

## Abstract

Due to the increasing popularity of microblogging platforms (e.g., Twitter), detecting real-time news from microblogs (e.g., tweets) has recently drawn a lot of attention. Most of the previous work on this subject detect news by analyzing propagation patterns of microblogs. This approach has two limitations: (i) many non-news microblogs (e.g. marketing activities) have propagation patterns similar to news microblogs and therefore they can be falsely reported as news; (ii) using propagation patterns to identify news involves a time delay until the pattern is formed, therefore news are not detected in real time. We propose an alternative approach, which, motivated by the necessity of real-time detection of news, does not rely on propagation of posts. Moreover, we propose a real-time sorting strategy that orders the detected news microblogs using a translational approach. An experimental evaluation on a large-scale microblogging dataset demonstrates the effectiveness of our approach.

## 1 Introduction

Microblogging platforms (e.g., Twitter or SinaWeibo) have become very popular and their role as news media has been recognized. As people actively talk about what is happening, microblogs are the place where the first-hand news appear. Actually, over 85% of the leading topics on Twitter are news by nature (H. P et al, 2010).

Most of the recent works on news detection from microblogs rely on using temporal patterns of propagation (G. L. et al, 2010; R. G. and K. Lerman, 2010; J. L and J. Yang, 2011; F. Z. et al, 2012). As an instance, (F. Z. et al, 2012) assumes that bursty topics in microblogs correspond to events that have attracted the most online attention. To find such events, this work uses a model to detect busty topics, assuming that a global event is likely to follow a time-dependent global topic distribution. Although detection methods relying on the propagation characteristics of microblogs are based on reasonable assumptions, they have certain limitations. First, some microblogs not related to news have very similar propagation characteristics as news microblogs. For example, a microblog with a promotion or a gift may follow a similar propagation pattern as a popular news event. Second, propagation behavior can only be analyzed after a microblog has been posted for a certain amount of time. Previous work on news detection based on propagation knowledge cannot perform real-time detection, since propagation knowledge can only be obtained a period of time after microblogs are published. Some works explicitly mention that trying to detect microblogs using propagation knowledge in a short time reduces the effectiveness. For example, in (G. L. et al, 2010), experiments on Twitter data show that using 1-day propagation knowledge can mainly detect topics related to daily activities; only by using a 2-days history this method can detect some real emerging topics.

Therefore, using propagation characteristics is not a good idea if the objective is to detect news as soon as possible. An additional challenge is how to sort and present the newborn news microblogs according to their importance, Most of the current news detection platforms sort the microblogs by their pub-



*29th Pacific Asia Conference on Language, Information and Computation pages 462 - 470*
*Shanghai, China, October 30 - November 1, 2015*

lication time or their popularity. However, at any point in time, there can be lots of newborn news microblogs all of which have close publication time; thus, sorting them by the publication time may fail to show important news on the top. Besides, as we mentioned before, newborn news microblogs have limited prorogation information; thus, it is very difficult to access the popularity of newborn news microblogs. In this paper, we propose an alternative system for detecting and sorting microblogs with news in real time. Our framework does not rely on any propagation knowledge. Our system consists of three modules: news-microblog expert detection, news microblog detection, and news sorting. We observe that there exists a group of expert users, whose microblogs are all of a single type (e.g., news). In the first module, we apply a methodology for selecting *expert users* based on their *professionalism* and *activity*. By simulating the training corpus as the microblogs by the experts, the second module builds an ensemble classifier to detect news microblogs. The ensemble model combines weak classifiers trained from the corpora of different experts into a strong classifier. Moreover, it can be updated with low cost: once an expert posts some news microblogs, we only need to update the module corresponding to its corpus instead of the whole model. The third module defines a score for each detected news microblog, in order to rank these microblogs. In this module, we firstly propose a novel text representation called *Behavior-Actor-Venue bag of words* (BAVbow) for news microblogs which consolidates the most informative text from them. Then, we apply *value transfer with confidence* on the BAVbow representation, using the scores of the training corpus to rank the new microblogs whose scores are unknown.

We conduct experiments on data obtained from the microblogging service SinaWeibo, one of the most popular sites in China, used by well over 30% of Chinese Internet users, with a similar market penetration as Twitter. The effectiveness of each module is verified based on information collected by a group of users.

The remainder of this paper is organized as follows. In Section 2, we introduce our methodology by discussing in detail the news detection framework and the three sub-modules. Section 3 presents our experimental analysis. We conclude the paper and suggest directions for possible future work in Section 4.

## 2 Our methodology

Our system includes three modules. In a *training session*, the *News-microblog Expert Detection* module detects a set of microblogging users who actively post news microblogs. The posts by these experts forms the training corpus of news microblogs, used to train the other two modules: the *Expert-ensemble Classifier* and the *BAV Sorter*. The *Expert-ensemble Classifier* (Section 2.2) is used to classify newborn microblogs to news or non-news. It combines base classifiers constructed from the experts' corpus by considering the *professionalism* and *activity* degrees of experts. The *BAV Sorter* module (Section 2.3) provides a new representation method for news microblogs and employs a value transfer strategy to define an importance score for each new post classified as news by the *Expert-ensemble Classifier*. After the system has been trained, the newly posted microblogs can be classified as news/non-news by the *Expert-ensemble Classifier*, and those posts detected to be news can be ranked according to their importance by the *BAV Sorter* module. In the remainder of this section, we describe in detail the three modules.

### 2.1 News-microblog Expert Detection

Since microblogging data are large and they are updated at a high rate, it is not possible to manually label them. As an alternative, we propose an automatic corpus construction method, motivated by the observation that there exists a group of users whose microblogs are of a single type only. In the news domain, some real-world examples include: *@头条新闻 (#breaking news#)* from SinaWeibo and *@BBCWorld* from Twitter, which always post news microblogs. Next, we present our methodology for finding out these users which we call *news experts*. The selection strategy considers two characteristics of users: *professionalism* and *activity*.

#### 2.1.1 Expert Candidates Retrieval via User Profile

Microblogging platforms have a very large number of users and it is impossible to analyze the microblogs written by all of them. Thus, it is necessary to select a subset of them, which is expected

to include the news experts. Search for news experts will then be confined to this subset. There are two types of data that describe a user: his/her profile and the microblogs he/she posts. Profile information can be divided into three parts: (i), Description: This part includes usernames and other descriptive data given by the users themselves. (ii), Authority: Microblogging platforms provide verifications for some users, called *verified accounts*. Verification is currently used to establish authenticity in Twitter. The verified badge helps users toward discovering high-quality sources of information. (iii), Influence: A natural feature that indicates the influence of a microblogging user is the number of followers, since this number indicates how many people are reading the user's posts. After analyzing a certain amount of users whose microblogs focus on news, we found that some discriminative characteristics exist in their profiles. First, their descriptions always contain some keywords related to the type of microblogs they focus on ("新闻" and news, in the examples). Second, all of them have verified accounts. Third, they have high influence, i.e., they have at least a certain number of followers.

Therefore, to retrieve candidates of news experts, we can first define some news-related keywords $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$ and obtain candidate users to be news experts as follows. For each $w_i \in W$, we define a set $e^c(w_i)$ by selecting from the set of users those who (i) have $w_i$ contained in their description, (ii) are verified, and (iii) have at least $\theta$ followers. Then, the set $EC$ of *expert candidates* is defined by $\mathcal{E}^c = \bigcup_{i=1}^{n} e^c(w_i)$. Note that both of the keyword set $W$ and $\theta$ can be seen as model parameters; they influence the number of retrieved expert candidates. For example, decreasing $\theta$ increases the number of candidates exponentially, therefore $\theta$ should be selected to be relatively large (e.g., we set $\theta$ =1,000 in our experiments). Still, as we will see next, since only the $k$ most important candidates will be selected in the end as experts, the overall training cost of our system is not very sensitive to these parameters.

#### 2.1.2 Selection of Professional and Active Experts among Candidates

After finding the set $\mathcal{E}^c$ of expert candidates, our method, as its second step, defines a score for each user in $\mathcal{E}^c$, which quantifies the candidate's appropriateness. The selection is based on the following rules: (i) microblogs posted by experts should focus on the type we are interested in (i.e., news), (ii) experts should be active, so that they provide time-relevant microblogs to be used in training our classification model. Thus, a candidate expert is more *professional* if a large percentage of his/her microblogs belong to the type. The more *active* the expert is, the more up-to-date his/her corpus is and the more adaptive it is to newborn microblogs.

Based on the above, we define the *professionalism* and *activity* degree of each candidate. To measure the professionalism of a candidate $e^c \in \mathcal{E}^C$, we need to use a classifier which indicates whether a post by $e^c$ is a piece of news. However, when the system is firstly used, we do not have a training set for such a classifier. To tackle this problem, we define a special corpus called *exterior-professional corpus*, which is not taken from microblogging platforms but from professional news sources, e.g., news web sites. We use the content of these sites to train a classifier $\mathbb{C}$ to evaluate the professionalism degree of news-expert candidates. Note that the resulting classifier is not expected to be very accurate, since it is based on a corpus that does not consist of microblogs. However, here we only need $\mathbb{C}$ to *rank* the candidates based on their *professionalism* and this classifier does a good job in this direction: more professional experts typically get higher classification accuracy. Another problem is that the exterior-professional corpus only contains positive instances, while to train a classifier, we usually also need negative instances. This issue could be alleviated by the use of one-class classification methods (M. Y. and L.M. Manevitz, 2002). As an alternative, in our case (i.e., news detection), we construct a corpus of non-news microblogs as follows: we randomly extract microblogs from users not in the candidate set $\mathcal{E}^c$ and use microblogs by them with short content and limited forwarding. These microblogs have low probability to be news microblogs.

For each expert candidate $e_i^c \in \mathcal{E}^c$, we extract recent posts (e.g., posted during time interval $[T_b, T_e]$, where $T_e$ is the extraction time ) by $e_i^c$ as $\mathcal{M}_i^T$. Then, we compute (i) $ec_i$'s professionalism degree using classifier $\mathbb{C}$: $f^p(e_i^c) = \frac{n'}{n}$, where $n$ is the number of posts in $\mathcal{M}_i^T$ and $n'$ is the number of posts in

$\mathcal{M}_i^T$ classified as news by $\mathbb{C}$; (ii) $e_i^c$'s activity degree as her posting frequency in a recent time interval (e.g., the last month): $f^a(e_i^c) = \frac{n}{T_e - max(T_b, T_u^i)}$, where $T_u^i$ is the time when $e_i^c$ registered at the microblogging platform. Then, the PA score $f^s(ec_i)$ is computed for the candidate as a weighted average of the user's normalized professionalism and activity (based on a weighing parameter $\alpha$). Finally, the $k$ users in $\mathcal{E}^c$ with the highest PA scores are selected as experts.

### 2.2 Expert-ensemble Classifier for Detecting News-Microblogs

After obtaining a set of news experts (denoted as $\mathcal{E}$) and their PA values, we utilize them to construct a classification model to identify whether a microblog is related to news. Considering this, we make use of the ensemble learning theory (Z. Zhou, 2012) to construct a classifier that detects news microblogs. First, given an expert $e_i \in \mathcal{E}$, we build a base classifier $\mathbb{C}_i$ corresponding to $e_i$'s corpus (i.e., $e_i$'s recent microblogs). To select an appropriate model for the base classifier in our experimental part, we first performed a comparison among the possible methods for training (not included in this paper, due to space constraints), and decided to use Multinomial Naive Bayes (MNB) (G. P. et al, 2006), owing to its good performance. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. MNB is one of the two classic Naive Bayes variants used in text classification, where the data are typically represented as vectors of word counts and tf-idf vectors. Note that other classification methods also can be used for constructing base classifiers. Users of our methodology can select by comparing performance of different classification methods in their own data.

After $k$ base classifiers $\mathbb{C}_1, \mathbb{C}_2, \cdots, \mathbb{C}_k$ are built (one for each of the $k$ experts), when a newborn microblog $m$ is posted, we can obtain a prediction from each $\mathbb{C}_i$ $(i = 1, \cdots, k)$ on whether $m$ is a news microblog. We use a indication function $f_i^{\mathbb{C}}$ to "binarize" the output of $\mathbb{C}_i$: If $\mathbb{C}_i$ classifies input $m$ as news, we set $f_i^{\mathbb{C}}(m) = 1$; otherwise $f_i^{\mathbb{C}}(m) = -1$.

Finally, by taking the PA values as weights, we aggregate the predictions of $k$ base classifiers to a single prediction in the ensemble manner:

$$f^N(m) = \sum_{i=1}^{k} f^S(e_i) \times f_i^{\mathbb{C}}(m). \quad (1)$$

A positive $f^N(m)$ value indicates that $m$ is a news microblog.

We use ensemble theory to construct the classifier for the following reasons. First, it matches the structure of input data (news experts' microblogs and PA values). Moreover, PA values are proportional to the confidence of the prediction based on each expert's corpus, since a more processional and active corpus derives a more accurate prediction in principle. Last, the ensemble can be updated at a low cost: the final model (Equation 1) is a linear combination of the sub-models (i.e., $\mathbb{C}_i$'s); each $\mathbb{C}_i$ is based on the corpus of a single news expert $ec_i$. Therefore, if one expert's corpus is updated (e.g., $e_i$ posts a number of new microblogs), only one sub-model (e.g., $\mathbb{C}_i$) needs to be updated.

### 2.3 BAV Sorting

The output of the ensemble classifier is just whether a newborn microblog is related to news. Therefore it is likely that a large number of newborn microblogs are classified as news. In order not to overwhelm the user of our system with a potentially huge number of news items, we can *rank* the items and present to the user only the most important ones. However, most of the research on ranking microblogs focus on subjective criteria (e.g., search-based (Y. Duan et al, 2010) or personalized ranking (W. C and I. Uysal, 2011)), which are not suitable for our problem (i.e., the detected newborn news are not based on search keywords or some user). Moreover, the ensemble prediction score (output of Equation 1) is proportional to the confidence of the predicted label, which may be independent to the importance of the classified news post. Therefore, we propose a novel ranking method for newborn news microblogs, called *Behavior-Actor-Venue based Sorting* (BAV sorting, for short). Firstly, for informatively representing news microblogs, we propose method called *Behavior-Actor-Venue BOW* ($BAV_{bow}$). Then, we apply a *value-transfer with confidence* method to transfer the knowledge of user-defined scores on a training set of microblogs, to the newborn microblogs, which can then be ranked based on their

predicted scores.

#### 2.3.1 Behavior-Actor-Venue Representation

Before text analysis, a common pre-processing approach *bag of words* (BOW)(J. L. et al, 2009) converts each document to a set of words, disregarding grammar and even word order. The BOW result is typically improved by eliminating uninformative or noisy terms. Recall that our work focuses on news microblogs; the most important components of a news item can be obtained by three questions: "Where it happened" (Venue),"Who did it" (Actor), and "Did what" (Behavior). Therefore, we design a representation model called *Behavior-Actor-Venue Bag-Of-Word* ($BAV_{bow}$) for analyzing news microblogs, which only keeps terms related to the information structure of news. These terms are extracted by making use of natural language processing (NLP) techniques, more specifically *Part-Of-Speech* (POS) tagging (S. J. DeRose, 1988) and *Named Entity Recognition* (NER) (F. Abedini et al, 2011). A common use of POS tagging is the identification of words as nouns, verbs, adjectives, adverbs, etc. NER locates and classifies atomic elements in text into predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. In news microblogs, verbs always indicate "Did what" (Behavior); persons and organizations recognized by NER always represent "Who did it" (Actors); "Where it happened" (Venues) can be found via Location terms (also recognized by NER). Therefore, in $BAV_{bow}$, only verbs and nouns related to people, organizations, or locations are kept. For example, a $BAV_{bow}$ representation of "In this morning, Lee had a big parade in Beijing." is {Lee, had, parade, Beijing}.

#### 2.3.2 Importance-Value Transfer with Confidence (VTC)

To assess the importance value $V^I(m)$ of a newborn microblog $m$ in a $BAV_{bow}$ representation, we make use of the concept of *knowledge transfer*; i.e., we transfer the knowledge about the importance values of old microblogs to the unknown values of newborn microblogs. There are several ways to define the importance values of old microblogs. One natural idea is employing the popularity (propagation characteristics) to calculate the importance value. For example, in the experimental part, we simulate the importance value of an old news microblog by how fast it spread among users of the platform. To model the *spread* of an item $m$, we use the number of reviews $n^r(m)$ and forwards $n^f(m)$ of $m$ in a certain time window after $m$ was published (e.g., one month):

$$V^I(m) = n^r(m) + \beta n^f(m), \qquad (2)$$

$\beta$ is a parameter to control the proportion of two terms.

An alternative way to implement this approach is to allow the users to rate the old news microblogs by themselves. Then the final order of news microblogs will conform to the subjective perception or interest of the users. For example, if they are more interested in news in some categories (e.g., technology), they would give higher values to the old microblogs in these categories; if they are more interested in some particular authors, they would value higher the old microblogs posted by or related to these authors.

Once we we have a *training* set of microblogs $\mathcal{M}_{tr}$ whose $Val$ values are already known (i.e., computed based on spread values or defined by real experts), we can employ VTC to predict the $Val$-based ranking of news microblogs $\mathcal{M}_{te}$ whose $Val$ values are unknown yet, i.e., $\mathcal{M}_{te}$ is a set of newborn microblogs classified as news by our ensemble classifier.

VTC compares the $BAV_{bow}$ representation $BAV_{bow}(m_j)$ of each $m_j \in M_{te}$ to the $BAV_{bow}$ representations of all $m_i \in M_{tr}$, in order to transfer the $Val$ values of microblogs in $\mathcal{M}_{tr}$ to $V^I(m_j)$, based on the similarity of the representations. The intuition behind the transfer strategy is shown by the following example. Consider two posts $m_1$ and $m_2$ with different but overlapping $BAV_{bow}$ representations:
$BAV_{bow}(m_1)$: Lee parade Beijing
$BAV_{bow}(m_2)$: Lee parade
Knowing that $m_2$ is important provides strong evidence that $m_1$ is at least somewhat important. However, knowing that $m_1$ is very important does not allow us to conclude that $m_2$ is, since the importance value of $m_1$ might also stem from Beijing. Thus, we can infer that, considering a microblog $m_i$ of $\mathcal{M}_{tr}$ and a microblog $m_j$ of $\mathcal{M}_{te}$, if the term set of $m_j$ contains a large proportion of terms in $m_i$, then

$V^I(m_i)$ is transferred to $V^I(m_j)$ with high confidence. We define the confidence value of transferring the value of $m_i$ to $m_j$ as follows:

$$V^C(m_i \rightarrow m_j) = \frac{|BAV_{bow}(m_i) \cap BAV_{bow}(m_j)|}{|BAV_{bow}(m_i)|}, \quad (3)$$

where $|\cdot|$ denotes the cardinality of the enclosed set.

Specifically, for each document $m_j \in M_{te}$, we consider each document $m_i \in M_{tr}$, compute the corresponding confidence $V^C(m_i \rightarrow m_j)$, and then transfer the $Val$ value of $m_i$, with $V^C(m_i \rightarrow m_j)$ as a weight, to $m_j$:

$$V^I(m_j) = \frac{\sum_i V^I(m_i) \times V^C(m_i \rightarrow m_j)}{|_{m_i \in \mathcal{M}_{tr}} V^C(m_i \rightarrow m_j) \neq 0|}; \quad (4)$$

The denominator of Equation 4 is the number of posts in $\mathcal{M}_{tr}$ for which the transfer confidence to $m_j$ is non-zero. The objective of VTC is to sort the newborn microblogs $m_j \in M_{te}$ in increasing order of their predicted importance values $V^I(m_j)$.

## 3 Experimental Evaluation

In this section, we evaluate the effectiveness our proposed system. To test our method on news mining from microblogging platforms, we applied it on a collection of microblogs with no propagation knowledge from the SinaWeibo platform. To assess the accuracy of our results, we invited 10 experts to provide correct labels (news vs non-news) to the tested microblogs. All these experts have journalism and linguistics background and their help is acknowledged at the end of the paper. In our evaluation, we divide 1,000 test examples of microblogs in 10 folds. The real labels (i.e., news/non-news) of each fold (containing 100 cases) are evaluated by the experts.

### 3.1 Experiments on News-Expert Retrieval

As introduced in Section 2.1, our system constructs a microblogging training corpus by selecting news experts. We first define some news related keywords $W$ ={新闻#news#,日报#dairly#, 时报#times#, 晨报, 晚报, 周报, #newspaper#} and then use them as queries to obtain news expert candidates with the threshold $\theta$ of minimum number of followers set to 1,000. For constructing the exterior professional corpus, we extracted 70,000 news titles for the period 2008/01/01-2012/12/31 from a news website[1]. Besides, as negative samples for our professionalism classifier $\mathbb{C}$ (see Section 2.1.2), we selected a non-news microblogging corpus of nrealtime 300,000 microblogs. Each has less than 30 Chinese characters and less than 50 forwards.

Finally, we obtained 486 expert candidates and their PA values. The parameter $\alpha$, which is used in combining professionalism and activity was set to 0.6, giving slightly higher weight to professionalism. Cross-validation (R.Kohavi et al, 2012) could be used to fine-tune this value. Table 1 indicatively shows the expert candidates with the top 3 and bottom 3 PA values. By analyzing their profiles and the microblogs posted by these candidates, we observed that most of these candidates have potential to be news experts since, compared to other users, their microblogs are more focused on news.[2] In order to verify the suitability of our ranking over these candidates (i.e., the suitability of the PA values obtained by our method), we estimated their real professionalism and activity as follows. For each candidate, we extracted the 10 most recently posted microblogs. The number of real news microblogs (as labeled by our invited evaluators) in the fragment (denoted as News_in_10) and the timespan of the microblog sequence (denoted as Time_for_10) are shown in the last two columns of Table 1. By looking at these results, we can see that the PA order indeed reflects the professionalism (i.e., high News_in_10 value) and activity (i.e., low Time_for_10 value) of the users.

After obtaining the expert candidates and their PA values, we select the $k$ candidates with the highest PA values to be the news experts. In order to determine $k$, we examine the PA values of the 467 candidates in our experiment in decreasing order. By looking closely at this sequence at the area around the 100th expert, we observe that there are several relatively sharp drops. In order to select the best $k$, we compute the moving average (the window size equals to 20) and compare it with the individual PA values. We select the PA value having the largest difference from the moving average value at that point. This value corresponds to the 109-th rank, thus we select $k$ =108.

[1] http://news.sina.com.cn/media.html

[2] This observation is verified by the invited evaluators familiar with Chinese media.

Table 1: Top-5 and Bottom-5 candidates of news experts

| Username | News_in_10 (number) | Time_for_10 (hours) |
|---|---|---|
| Top 5 | | |
| 头条新闻#Breaking News# | 10 | 5 |
| 法制日报#Legal Daily# | 10 | 33 |
| 新闻晨报#Morning Post# | 8 | 3 |
| 西安新闻网#Xi'an Web News# | 9 | 26 |
| 宁波日报#Ningbo Daily# | 8 | 16 |
| Bottom 5 | | |
| 每日甘肃网#Gansu Web Daily# | 8 | 71 |
| 江西五套#Jiangxi Channel5# | 8 | 330 |
| 晨报周刊#Morning Post Weekly# | 5 | 31 |
| 光明日报#Guangming Daily# | 7 | 7 |
| 辽沈晚报大活动#LiaoShen Evening Activities# | 3 | 188 |

### 3.2 Experiments on News-microblog Classification

To perform detection of news microblogs, we used the posted microblogs (scale: 1,335,884 microblogs) of the $k$=108 news experts in the period 01/01/2012 to 31/12/2012. The test set (scale: 610,000 microblogs) was collected during January 2013 from about 32,000 users, randomly selected from the whole set of SinaWeibo users. We trained our model (denoted as *Expert-ensemble_pa*) to classify the test microblogs. Here, we compare our method with a natural method as the baseline (denoted as *Single*), which is used in TwitterStand (J. Sankaranarayanan et al, 2009): combine the microblogs from all news experts into a single corpus and train a single classifier. Moreover, to evaluate the fitness of using PA values to be ensemble weights, we also compared our method with Majority Voting (T. G. Dietterich, 2000) (denoted as *Expert-ensemble_mv*) which is a popular method in ensemble learning.

Since the scale of test microblogs is too large, we randomly selected 1,000 of them (including 400 predicted news results and 600 predicted non-news results) and asked our evaluators to label them. Based on the labeling and the prediction by the classifiers, we derived the average performance of the classifiers as shown in Table 2. As Table 2 indicates, our method outperforms the *Single* classifier in all evaluation terms, especially in the precision of news category. This indicates that although the *Single* method can extract most of the news microblogs (i.e. not a bad recall on news), the predicted news microblogs are mixed with a significant number of non-news microblogs. On the other hand, our method uses the PA values of experts as confidences of the individual classifiers in the ensemble and this gives a large improvement in the accuracy of news microblogs classification. Another observation is that the accuracy in predicting non-news is higher than that of predicting news. In other words, the probability of mistaking a news item as a non-news microblog is lower than taking a non-news item as news microblog. Since the cost of mistaking news as non-news microblogs is higher, this result can be considered good for real-world applications.

### 3.3 Experiments on BAV Sorting

Our method is independent of the definition of importance value $Val$ for news, which may vary in different applications. In this experiment, we use as $Val$ the spread range (as defined by Eq. 2 with $\beta = 4$ and a one-month time window) of microblogs written by news experts during 01/01/2012 to 31/12/2012. The test microblogs are taken from detected news microblogs from the classification experiment (Section 3.2).

Our sorting method is based on (i) the BAVbow representation and (ii) the Value Transfer with Confidence (VTC) approach. To evaluate the effectiveness of using both BAVbow and VTC (BAVbow + VTC) and using BOW and VTC (BOW + VTC), with a random ordering of news (Random). The comparison between BAVbow + VTC and BOW + VTC shows the effectiveness of our proposed BAVbow representation. The comparison between

Table 2: Classification performance

| Evaluation terms | Single | Expert-ensemble_mv | Expert-ensemble_pa |
|---|---|---|---|
| Precsion on News | 72.6% | 86.3% | 92.3% |
| Precsion on Non-news | 95.1% | 85.8% | 96.6% |
| Recall on News | 88.9% | 95.1% | 93.8% |
| Recall on Non-news | 86.6% | 94.2% | 95.9% |
| Overall Precsion | 87.2% | 92.1% | 95.1% |

BAVbow (or BOW) + VTC and Random shows the effectiveness of VTC on predicting the importance values of news. For reducing the contingency of Random, we perform the comparison on 10 folds (each fold has 100 test microblogs) and average the results. For exploring the relationship between the real order and the order predicted by our method, in Figure 1, we show the average predicted rank sequences (BAVbow (or BOW) + VTC rank) as a function of the real ranked sequence $(1 - 100)$. To assess the relationship between the real order[3] and the predicted order, we also plot the smoothed lines (by minimizing the least squares error) for BAVbow (or BOW) + VTC. Besides, we include lines corresponding to a Random order and the perfect order prediction.

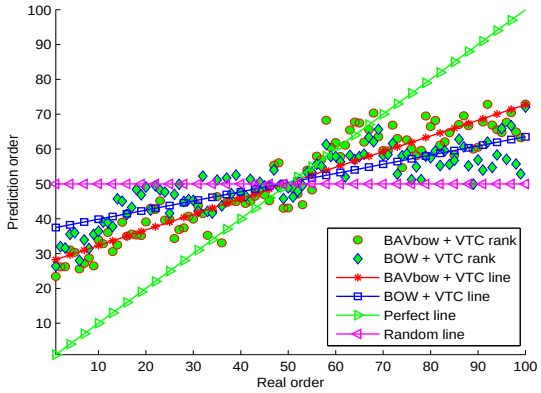


Figure 1: Comparison among BAVbow + VTC, BOW + VTC, perfect and random predictions.

As shown in the Figure 1, BAVbow + VTC is better than BOW + VTC, and the BAVbow + VTC line is closer to the perfect line, which indicates a good performance of our proposed BAVbow on representing news microblogs for knowledge transfer. Moreover, both of BAVbow + VTC and BOW + VTC are positively correlated with the correct ranking, as opposed to the Random line. We note that we also tested a method, which ranks the news posts according to the output of Equation 1 (i.e., confidence of the ensemble classifier) and found that its effectiveness is similar to that of the Random ordering.

[3]We compute the real values of formulation (2) of the test microblogs and then rank them to obtain the real order.

## 4 Conclusion

We have proposed a methodology for real-time news detection and sorting from microblogging platforms. Our approach automatically selects a set of users who are microblogging news experts. Based on the microblogs already posted by these expert users, together with the professionalism and activity knowledge of experts, we build an expert-ensemble classifier for detecting news microblogs. Then, newborn microblogs, which do not have any prorogation knowledge, will be classified as news or non-news ones. Going one step further, we propose a BAVbow + VTC sorting approach, which orders the detected news microblogs based on their expected value.

## References

F. Abedini, F. Mahmoudi, and A. Jadidinejad. From text to knowledge: Semantic entity extraction using yago ontology. *International Journal of Machine Learning and Computing*, 1(2),2011.

F. Z., E. L., Q. Diao, and J. Jiang. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536–544. Association for Computational Linguistics, 2012.

G. L., C. S., M. Cataldi, and C. Di. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

G. P., V. Metsis, and I. Androutsopoulos. Spam filtering with naive bayes-which naive bayes. In *Proceedings of the 3rd Conference on Email and Anti-spam*, volume 17, pages 28–69, 2006.

H. P., S. M., H. Kwak, and C. Lee. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

J. L., K. Weinberger, A. Dasgupta et al. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.

J. L. and J. Yang. Patterns of temporal variation in online media. In *Proceedings of the 4th ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.

M. Y. and L.M. Manevitz. One-class svms for document classification. *The Journal of Machine Learning Research*, 2:139–154, 2002.

R. G. and K. Lerman. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media*, 2010.

R.Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Articial Intelligence*, volume 14, pages 1137–1145, 1995.

S. J. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39, 1988.

T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 2000.

W. C. and I. Uysal. User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2261–2264. ACM, 2011.

Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010.

Z. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall, 2012.