2-2016

Ranking of high-value social audiences on Twitter

Siaw Ling LO
Singapore Management University, sllo@smu.edu.sg

Raymond CHIONG

David CORNFORTH

Ranking of High-Value Social Audiences on Twitter

Siaw Ling LO*, Raymond CHIONG, David CORNFORTH

School of Design, Communication and Information Technology, The University of Newcastle, Callaghan, NSW 2308, Australia.

*Corresponding author.

E-mail addresses: siawling.lo@uon.edu.au (Siaw Ling LO), raymond.chiong@newcastle.edu.au (Raymond CHIONG), david.cornforth@newcastle.edu.au (David CORNFORTH)

Abstract

Even though social media offers plenty of business opportunities, for a company to identify the right audience from the massive amount of social media data is highly challenging given finite resources and marketing budgets. In this paper, we present a ranking mechanism that is capable of identifying the top-k social audience members on Twitter based on an index. Data from three different Twitter business account owners was used in our experiments to validate this ranking mechanism. The results show that the index developed using a combination of semi-supervised and supervised learning methods is indeed generic enough to retrieve relevant audience members from the three different datasets. This approach of combining Fuzzy Match, Twitter Latent Dirichlet Allocation and Support Vector Machine Ensemble is able to leverage on the content of account owners to construct seed words and training datasets with minimal annotation efforts. We conclude that this ranking mechanism has the potential to be adopted in real-world applications for differentiating prospective customers from the general audience and enabling market segmentation for better business decision-making.

Keywords

Ranking, Audience segmentation, Social audience, Ensemble learning, Twitter

1. Introduction

In this age of information overload, the ability to identify relevant content in a timely manner will help both consumers and business entities in their decision-making processes. This is especially so when a company wants to find potential customers or a target audience from the crowded social media space. While most companies have Twitter or Facebook accounts [1], it remains a challenge for

them to fully leverage on the content shared on social media platforms in order to gain business insights or improve customer engagement.

Due to the privacy policy of Facebook profiles, this work focuses on Twitter, where most of the content and activities shared online are open and available. Twitter allows its registered users or account owners to send and read short messages (up to 140 characters) called *tweets*. Twitter users may subscribe to or follow other users' tweets and thus, the subscribers are also known as *followers*. While it is logical to assume that most followers who subscribe to a particular account would be interested in the content shared by the account owner, it is not uncommon to see followers subscribing to accounts they have no interest in through a marketing campaign or free product sample offer. A mechanism for the account owner to distinguish or classify followers who are genuinely interested in the content shared is therefore highly desirable, so that appropriate offers can be effectively sent to the right audience.

In order to help a company to understand their social audience and have the ability to segment them according to its business focus, we propose an approach to identify users' interests and rank selected users so that a practical solution can be offered to the company. Our proposed approach is capable of identifying a group of relevant followers and ranking them to help a company zoom into a segment of its online audience that most likely would be interested in the current business plan and hence equipping the company to devise strategies to better engage the online audience. This segment of online audience can be termed as the *high-value social audience* (HVSA). This HVSA is different from a group of influencers, since the latter usually consists of one or two persons who are authoritative in their domain but may or may not be a follower of the account owner. While success stories of using influencers can be found in social media campaigns [2], it can also backfire when an intended idea is not perceived well [3].

There are various methods for identifying an online target audience through shared content or platform-related features (see Section 2 for details). The focus is typically on classifying a group of users based on some specific interests [4][5] or categorising them into different demographics by a segmentation process [6][7]. Recently, some researchers have used a ranking approach [8] to discover top-k target users for advertising on Digg [9]. Another group of researchers have proposed a different

ranking approach for identifying suitable Twitter users to target when posting tweets [10]. To the best of our knowledge, none of the current approaches has attempted to identify a HVSA from the list of followers of an account owner and rank the potential social audience members so that marketing dollars can be used more effectively. We therefore believe that an approach like the one we propose in this paper is necessary, and that it will provide a better decision-making mechanism for companies engaging in social media, allowing them to be equipped with an ability to filter groups of social audiences depending on available resources or marketing budgets.

To rank the followers, we derive a HVSA index based on some of the scoring schemas previously developed in [11]. In addition, we make use of various evaluation metrics (e.g., Precision@k, Average Precision@k) common to the Information Retrieval (IR) domain, where significant efforts have been made to evaluate the top documents retrieved [12]. We also introduce a pooling strategy to extract an unseen set of testing data on top of the de-facto annotated testing dataset for assessing the scoring schemas and indices derived.

For evaluation purposes, three subjects or business companies of different nature have been selected. They are *Samsung Singapore* (samsungsg), *I Love Deals Singapore* (ilovedealssg) and *Be Aqua Fitness* (beaquafitness). For each of the datasets, we use content from the account owner as the positive training dataset while a negative dataset is constructed using data of other account owners based on domains discovered from followers via a semi-supervised topic modelling approach, i.e., Twitter Latent Dirichlet Allocation (LDA) [13]. This approach is able to create an annotated training dataset with minimum annotation effort for the subsequent machine learning process. Various methods including Fuzzy Match, Twitter LDA and Support Vector Machine (SVM) Ensembles [14] with different feature sets are assessed for their ability to identify and rank the HVSA. Four scoring schemas for representing Twitter users are used in this study and two HVSA indices, which show potential in ranking the HVSA regardless of the nature of the Twitter account owner's dataset, are investigated.

The main contributions of this work can be summarised as follows:

- We define an approach to identify and rank the HVSA of a Twitter account owner with minimum annotation effort.

- To the best of our knowledge, our work in this paper is the first attempt to rank a social audience via an HVSA index from the list of followers on Twitter using a combination of semi-supervised and supervised learning methods that is capable of identifying the top-k HVSA members from three datasets of different nature.
- From the observation of our results, our proposed pooling strategy is capable of evaluating the ranking capability of various methods with minimum influence from the differences in datasets. Moreover, we conclude that the Average Precision@k evaluation method should be used instead of Precision@k as the former offers a more sensitive measurement for HVSA ranking.
- Audience segmentation based on Twitter LDA on top of the ranked HVSA empowers a company to make better decisions in customer engagement and personalised service.

2. Related Work

Even though tweets can be a rich source of information, the huge volume and real-time nature of tweets can sometimes result in noisy posting about daily lives. Being able to extract relevant information from tweets for user profiling is hence essential. The majority of existing approaches for classifying and identifying Twitter users are based on the use of textual features (e.g., content of tweets) [4][5][15] or platform related features (e.g., retweet features, social media network structure or user profile information) [16][6][7] .

Pennacchiotti and Popescu [4], for example, used machine learning and LDA [17] to analyse the content of users and various other features such as account profiles and tweeting behaviour in order to classify a user for three tasks with different characteristics: political affiliation detection, ethnicity identification and business affinity detection. It was observed that different features play different roles in identifying the preference of a specific business, detecting ethnicity or political affiliation. Interestingly, the results from LDA-based features have been consistently reliable across all tasks.

Yang et al. [5] looked at the temporal effect of Twitter content for classifying user interests in sports and politics. Instead of using tweets directly, they derived temporal information from word usage within the streams to boost classification accuracy of the SVM and Naïve Bayes. Both binary

and multi-class classifications were considered, and their approach was found to substantially outperform other methods in comparison. Michelson and Macskassy [15] presented work in discovering topics of interest by examining entities in tweets. They developed a "topic profile" to characterise users, and utilised Wikipedia as the knowledge base for entity disambiguation to determine high-level categories defined by these entities.

Encouraged by these promising findings, in particular the consistently good performance of LDA [4] and classification accuracy of machine learning [5], our proposed approach uses Twitter LDA and a SVM ensemble together with Fuzzy Match to identify and rank a target audience from a list of followers using content shared by a Twitter account owner without the need of any external knowledge base. In addition, we use tweets from the same temporal period to enable the analysis of specialised terms or new technology that may not have been updated in external knowledge bases but these entities can mostly be found in the tweets of the account owner. This makes the proposed approach more robust and able to perform well across various domains.

Other related studies include that of Hong et al. [16], who explored users' interests and behaviours by using the retweet action in Twitter to model user decisions and user-generated content simultaneously. Rao et al. [6] adopted various sociolinguistic features such as emoticons and character repetitions, and they used the SVM to classify latent attributes such as gender, age, regional origin and political orientation. Ikeda et al. [7] proposed some demographic estimation algorithms for profiling Japanese Twitter users based on their tweets and community relationships, where characteristic biases in the demographic segments of users were detected by clustering their followers and followees. As the aim of our study is to identify the HVSA, we focus on content shared by followers, which includes retweeted content, although we do not specifically consider the retweet action. While sociolinguistic features and demographic information are important for targeting potential customers, we have decided to concentrate on developing an approach that is able to identify followers who are more likely be interested in the content shared before expanding our study to sentiment analysis and demographics clustering.

Zhang and Pennacchiotti [18] showed that it is possible to predict e-commerce purchasing behaviour by using Facebook and related eBay data. However, as Facebook data is restricted by its

privacy policy and the eBay purchase dataset is not readily available, their approach is hard to replicate. Another study relying on Facebook data by van Dam and van de Velden [19] utilised Facebook users' "like"s for online profiling and clustering so that different segments can be identified through analysis of its Facebook fans. This work is similar to ours in the aspect of analysing strategic segmentation of social media users associated to a company, as we are analysing followers of Twitter account owners representing business companies. However, since there is no explicit "like" field for Twitter users to indicate their interest, we extract relevant entities from tweets and use various methods to identify followers with similar interests according to the content shared by the account owner.

While various approaches and features have been proposed to identify or classify a social audience, none of the previous studies had investigated the possibility of ranking the target audience so that segmentation can be done more effectively. Two recent studies that are closely related to our work can be found in Rao et al. [8] and Tang et al. [10]. Rao et al. [8] analysed a Digg dataset and identified the top-k most desirable target users who most likely will view the advertised information and perform potential e-commerce activities. They used Digg's content as well as social, location and time based features on a learning-to-rank framework, Ranked SVM [20], for the task. Tang et al. [10] proposed a ranking based recommendation approach based on *Twitter mentions* to identify top-k targets for advertising. Similarities between neighbouring users, estimated via content similarity and structural similarity, were used for measurement. Top-k query algorithms such as the Threshold Algorithm [21] and Not Random Access algorithm [22] were then applied to retrieve the list of target users. Our proposed approach is different to theirs as we have a generic scoring mechanism through a HVSA index derived from a combination of semi-supervised and supervised learning methods that is able to overcome the diversity and variance introduced in different datasets. In addition, our aim is to propose a ranking model for the social audience so that a company with online presence can use this approach directly on its social media followers (e.g., followers on Twitter) for marketing activities or to revise its online engagement plan.

3. Details of Datasets

For evaluation and comparison purposes, datasets from three Twitter account owners of different nature were used in this study. The first dataset was "samsungsg" (the official Twitter account for Samsung Singapore), the second dataset was "ilovedealssg" (a Twitter account for daily deals, promotions and discounts in Singapore), and the third dataset was "beaquafitness" (the Twitter account of a company focusing on aqua fitness solutions in South East Asia). As samsungsg is a technology, mobile, and appliances company, the content shared by it tends to be quite homogeneous. On the other hand, ilovedealssg often shares deals from multiple domains, and hence the content is heterogeneous. Being a company in aqua fitness, beaquafitness uses many specialised terms (e.g., fitness equipment) but it also touches on a variety of topics such as healthy living and training activities in its content.

3.1 Analysis of Datasets

To better understand the contents shared by the three account owners, OpenCalais [23], an open service that categorises text into multiple topics, was used to discover the type of topics belonging to the three accounts and the results can be found in Table 1. The numbers shown after the topics are raw counts of topics identified.

Table 1. OpenCalais results for the three Twitter account owners (samsungsg, ilovedealssg and beaquafitness)

Account owner (tweet analysed) [returned topic size]	samsungsg (199) [topic size 434]	ilovedealssg (184) [topic size 308]	beaquafitness (143) [topic size 232]
Top categories	Technology_Internet 70 Entertainment_Culture 50 Business_Finance 25 Hospitality_Recreation 18 Human Interest 18 Sports 18 Law_Crime 17 Politics 5 Disaster_Accident 4 Education 3 Religion_Belief 3 Health_Medical_Pharma 2 Social Issues 1 Labor 1	Hospitality_Recreation 53 Entertainment_Culture 14 Technology_Internet 14 Law_Crime 12 Human Interest 10 Business_Finance 7 Sports 5 Health_Medical_Pharma 4 Disaster_Accident 3 Religion_Belief 2 Environment 1	Hospitality_Recreation 79 Entertainment_Culture 25 Human Interest 24 Technology_Internet 19 Business_Finance 17 Health_Medical_Pharma 15 Sports 12 Education 11 Politics 7 Environment 6 Religion_Belief 5 Weather 4 Social Issues 3 Disaster_Accident 3 Labor 1

As can be seen in Table 1, samsungsg is dominated by two main topics while ilovedealssg has a mixed range of topics but the primary one is from the hospitality or recreation category. It is interesting to observe that the top two topics for beaquafitness are the same as ilovedealssg but with Hospitality_Recreation the dominating one. Although the business of beaquafitness is focusing on quite a distinctive domain, i.e., aqua fitness, its range of topics can be rather diverse.

3.2 Dataset Collection

Twitter's Search API was used for our data collection. As the API is constantly evolving with different rate limiting settings, our data gathering was done through a scheduled program that requests a set of data for a given query. 200 tweets of each account owner and the past 100 tweets of their followers of the same period were extracted. We chose to analyse only tweets from active followers or Twitter users who had shared five or more tweets during the specified period. Details of the datasets can be found in Table 2.

Table 2. Volume and period of datasets

Account owner	Number of followers (number of tweets)	Active followers (number of tweets)*	Period
samsungsg	3727 (187,746)	2449 (124,462)	2 Nov 2012 to 3 Apr 2013
ilovedealssg	1260 (58,880)	844 (57,114)	26 Mar 2013 to 15 Jul 2013
beaquafitness	179 (11,983)	143 (11,969)	05 Jan 2013 to 11 Nov 2015

*Followers who shared five or more tweets during the specified period.

3.3 Construction of Testing Datasets

While the training datasets were constructed with minimum manual annotation through semi-supervised learning of Twitter LDA, manual annotation was used for deriving testing datasets from active followers' tweets. Figure 1 shows how the testing datasets are constructed from followers' tweets. Besides randomly selecting a portion or roughly 10% of the size of the original followers dataset as an annotated follower (AF) dataset for validating results of the classifiers shown in Figure 2, a smaller number of annotated tweets (AT) are arbitrary chosen for the purpose of deriving a threshold for the count scoring schema of a Twitter user (see Section 5.1 for details). Essentially, the followers' tweets are segregated to three portions for different purposes to ensure that unseen data or feature sets are used for the various classification and evaluation processes.

In this study, 300, 100 and 50 followers were randomly selected as the AF testing datasets for samsungsg, ilovedealssg and beaquafitness, respectively. 2500, 1500 and 750 tweets were arbitrarily chosen as ATs for deriving the threshold value for the various SVM ensembles.

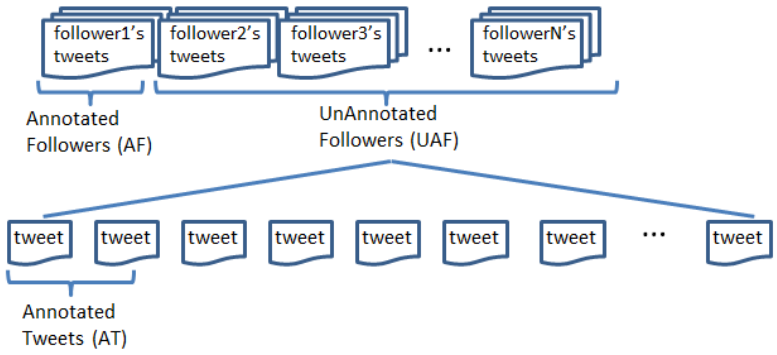

Figure 1. Construction of various testing datasets

4. HVSA Identification

In this section, we present the details of methods used for identifying the HVSA from followers of selected Twitter account owners - samsungsg, ilovedealssg and beaquafitness - with minimum annotation effort. Two of the methods, Fuzzy Match and Twitter LDA, have provided promising results in our preliminary study [11] based on the samsungsg dataset. The third, which combines semi-supervised (Twitter LDA) and supervised (SVM ensembles) learning, has shown to be able to differentiate a target audience from the general audience [24]. A simplified overall architecture diagram can be found in Figure 2.

As can be seen from the figure, a selected account owner's tweets are used as the positive training dataset while tweets from other account owners are used as the negative training dataset. The domains of other account owners are determined by studying their followers' contents using Twitter LDA before representative ones are chosen. This approach is capable of constructing a representative training dataset with minimal annotation effort. All tweets are first pre-processed to identify the relevant entities or phrases. A list of seed words are derived using content from the selected account owner so that Fuzzy Match and Twitter LDA analysis can be done using an annotated testing dataset. The results are then measured using *s* and *t* scores (described in Sections 4.1 and 4.2). Bagging and bootstrapping SVM ensembles with different feature representations are used to develop models for

classifying the annotated testing dataset, and a *v* score (described in Sections 4.3) is generated for each follower. Details of followers' domains discovery, seed words generation and pre-processing procedures can be found in [24].

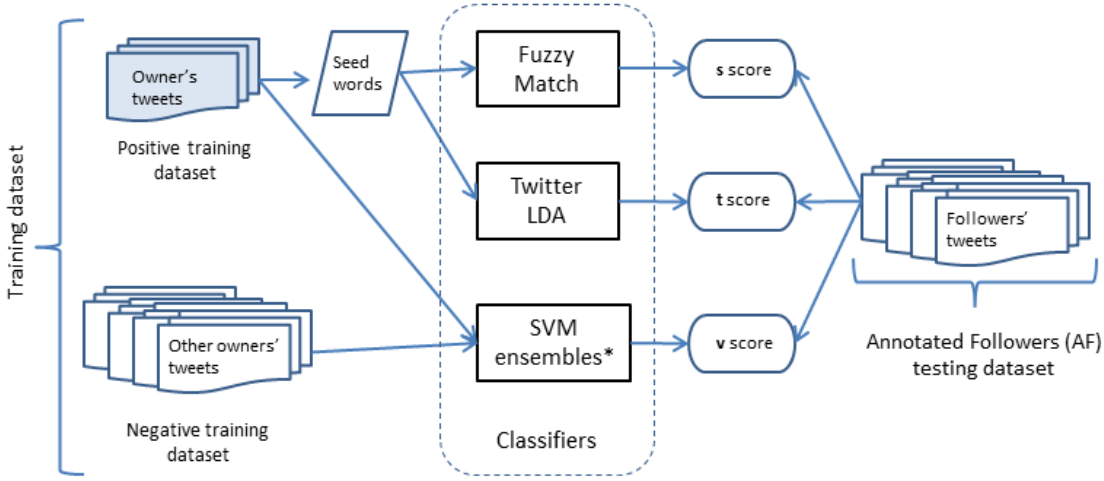

Figure 2. A simplified overall architecture for HVSA identification. *Various models with different feature representations (see Sections 4.3 and 5.1 for details).

4.1 Fuzzy Match

It is not uncommon for Twitter users to use abbreviations, interjections or different forms of expression to represent similar terms. For example, Samsung's phone product "galaxy s iii" can be represented by "galaxy s 3", which is understandable by a human but cannot be captured if a direct keyword match method is used. A Fuzzy Match method using seed words derived from the account owner's tweets is therefore a better option.

The comparison in this regard is based on a Dice coefficient string similarity score [25], which can be calculated as follows:

$$s = 2 \times n_t / (n_x + n_y) \qquad (1)$$

where n_t is the number of characters found in the strings to be compared (e.g., strings x and y), n_x is the number of characters in string x and n_y is the number of characters in string y. For instance, consider the calculation of similarity between "process" and "proceed":

x = process bigrams for x = {pr ro oc ce es ss}

y = proceed bigrams for y = {pr ro oc ce ee ed}

Both x and y have 6 bigrams each, of which 4 of them are the same. Hence, the Dice coefficient string similarity score or *s* score is 2*4/(6+6) = 0.67. Each tweet of a follower is compared with the seed words in this manner and the highest score of any match is maintained as the *s* score of the follower.

4.2 Twitter LDA

LDA [17], a renowned generative probabilistic model for topic discovery, has been used in various social media studies (e.g., see [13][26]). LDA uses an iterative process to build and refine a probabilistic model of documents, each containing a mixture of topics. However, standard LDA may not work well with Twitter, as tweets are typically very short. If one aggregates all the tweets of a follower to increase the size of the documents, this may diminish the fact that each tweet is usually about a single topic. Moreover, our previous study has shown that it is essential to represent each individual tweet as a single topic, as combining all the tweets to extract representative topics do not perform well in the context of SVM classification [11]. We therefore have adopted the implementation of Twitter LDA [13] for semi-supervised topic discovery.

Our previous work has also shown that a 20-topic Twitter LDA model performs better than 10- and 30-topic models in identifying HVSA [27]. For this reason, Twitter LDA with 20 topics was used in this study. We generated a list of 20 topics after running 100 iterations of Gibbs sampling while keeping the other model parameters (Dirichlet priors) constant: α = 0.5, β_{word} = 0.01, $\beta_{background}$ = 0.01 and γ = 20. Suitable topics were chosen automatically via comparison with the list of seed words. As Twitter LDA is an unsupervised learning approach, 30 runs were conducted to consolidate the topic assignment for each follower. In the end, a *t* score was assigned to each follower using the following calculation:

$$t = n_m / n_r \qquad (2)$$

where n_m is the total number of matches and n_r is the total number of runs. If any of the suitable topic was found in five runs for a particular follower (out of the 30 runs), the *t* score of that follower will be assigned as 5/30 = 0.17.

In addition to generating the *t* score for a follower, we also used Twitter LDA for audience segmentation on the list of top-k followers discovered from a pooling strategy (see Section 7.5.2). A

10-topic model was used for this purpose given the fact that the size of the dataset is relatively small. Topics of interest were selected through analysing of the top topical words and the corresponding topic ID was used to identify followers falling under the topic. Top five topics assigned to each follower were then analysed and compared to the selected topics of interest so that specific followers can be extracted.

4.3 SVM Ensembles

The SVM is a well-known supervised learning approach for two- or multi-class classification, and has been used successfully in text categorisation [14]. It separates a given known set of {+1, -1} labelled training data via a hyperplane that is maximally distant from the positive and negative samples. This optimally separating hyperplane in the feature space corresponds to a nonlinear decision boundary in the input space. More details of the SVM can be found in [28].

The LibSVM implementation of RapidMiner [29] was used in this study and the sigmoid kernel type was selected, since it produces higher precision prediction than other kernels, such as the Radial Basis Function and polynomial kernels. Based on the SVM, a *v* score was assigned to each follower according to individual tweet classification. The v_p score was generated using the following equation:

$$v_p = n_p / n_a \qquad (3)$$

where n_p is the total number of tweets that are classified as positive and n_a is the total number of tweets shared by a follower. Note that this score is also termed as the percentage *v* score, hence the subscript *p*. Here, we used the total number of tweets to normalise the score instead of an average value of all tweets. This way, the resulted score would be more capable of representing the true interest of a follower. For example, if follower1 tweeted two related tweets out of a total of 10 tweets, the v_p score assigned is 0.2, while the v_p score for follower2 is 0.02 if only two related tweets were classified as positive out of a total of 100 tweets. This is in contrast to using an average value, as both follower1 and follower2 would be assigned the same v_p score that may not fully represent the follower's interest.

As can be seen in Figure 2, there is a data imbalance issue introduced through the use of content from account owners. That is, negative training datasets that are formed by using data from other

account owners representing the followers' domains are often several times larger than the positive training dataset. To overcome this issue, we propose two types of SVM ensembles to leverage the diversity and also to ensure that there is no information loss in the majority class or introduction of bias in favour of the minority class. The first is a bootstrapping ensemble using a single SVM model, while the second is a bagging ensemble using multiple SVM models. These two ensembles have been shown to perform well [24]. The bootstrapping ensemble system has achieved the best result under 10-fold cross-validation while the bagging ensemble system performs best in classifying annotated unseen testing datasets. Figure 3 shows the construction of these two SVM ensembles in our case.

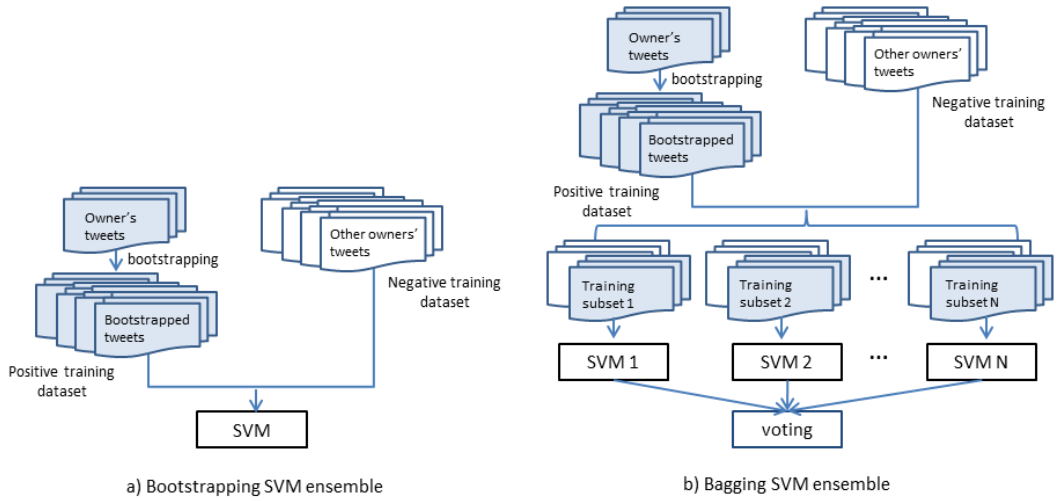

Figure 3. Construction of SVM ensembles

In order to derive a feature space for learning, a vector space needs to be created where each tweet is represented as a vector composed of words or phrases. In addition to Term Frequency (TF) used in [24], we also used Term Frequency Inverse Document Frequency (TFIDF) in this study to assess the influence of different weighting schemas of the SVM ensembles. As a result, four SVM ensembles were constructed for each account owner's dataset: bootstrapping SVM ensemble using TF (BT_TF), bootstrapping SVM ensemble using TFIDF (BT_TFIDF), bagging SVM ensemble using TF (BG_TF), and bagging SVM ensemble using TFIDF (BG_TFIDF).

5 HVSA Ranking

Apart from the *s*, *t* and v_p scores discussed in the previous section, we further propose three other scoring schemas for ranking the HVSA, which will be described in this section.

5.1 The Count Scoring Schema

Although the classification of SVM ensemble is through analysing each tweet shared by a Twitter user (either an account owner or a follower), our main purpose is to determine whether a follower is a HVSA. Hence, there is a need to derive a method that can represent a Twitter user through the tweets shared. The results from our previous study [11] have shown that it is essential to represent a Twitter user using individual tweets rather than 'summarise' the tweets into a single feature set. In this paper, we assess two possible ways (schemas) of representing a user taking into consideration all the tweets they shared:

1) Based on the percentage of number of tweets that are classified as positive and the v_p score is generated as per equation (3).
2) Based on the number of correct classifications from all the tweets shared and calculate the v_c score through a threshold.

As different SVM ensembles behave differently, there is a need to determine the suitable threshold to be used as a cut-off point in identifying a HVSA. It is important to highlight that the selection of a suitable threshold is a trade-off between the number of HVSA identified and the accuracy of selecting the 'true' audience who is highly likely interested in the content shared by an account owner. The algorithm used for this purpose is depicted in Figure 4.

Input:
D: Training dataset as shown in Figure 2
T: Testing dataset from UnAnnotated Followers (UAF) as shown in Figure 1
S: SVM ensemble learning algorithm
R: Integer specifying the number of records needed
O: Account owner

Do for each O, o
  Do for each S, s
    1. Generate s model from D_o
    2. Classify T_o on s model
    3. Based on the probability score generated, randomly select R_o from the correct class
    4. Manually annotate by assigning 1 to the correctly classified and 0 to the wrongly classified
    5. Calculate Youden's index [30] or threshold and the Area Under Curve (AUC) using probability scores and annotated labels
    6. Identify the probability score associated with the maximum Youden's index
    7. Assign the probability score to $Threshold_{O,S}$
  End
End

Figure 4. The threshold generation algorithm

Youden's index is defined as

$$Youden's\ index = Sensitivity + Specificity - 1 \quad (4)$$

where *Sensitivity* and *Specificity* are calculated for each point of the testing dataset (or each value of the probability score predicted by the classifier), and the point that generates the maximum (*Sensitivity* + *Specificity*) is the same as the maximum Youden's index when a single threshold value is required.

Even though Youden's index was originally used to capture the performance of a diagnostic test, where it is essentially the height measured above the 'chance' line in a Receiver Operating Characteristic (ROC) curve, it is applicable in this study as we want to find a cut-off point or threshold value that can maximise the sensitivity and specificity of the testing dataset. As this index is defined for every point on a ROC curve, the maximum value of the index can be used as a criterion for selecting the optimal threshold point.

After deciding on a threshold value for each SVM ensemble, probability scores from the AF testing dataset can then be processed to generate the v_c score. For each follower, we take the absolute count of tweets with a probability value greater or equal to the threshold decided. If the absolute count is the same for more than one follower, the average of correctly classified probability scores is added to each of the counts so that ranking can be done. For example, if two followers have two tweets with probability scores above the threshold, both of them will be assigned a v_c score of 2. However, as there is a need to rank them, the probability scores of all their correctly classified tweets will be averaged to add on to the count for deciding the final v_c score. This approach is to ensure that the follower who tweets more relevant content is ranked higher than those with tweets of lesser probability scores.

Given the different scoring schemas to represent a follower in the vector space, there are now eight SVM ensembles in total to encode the dataset of each account owner: BT_TF_C indicates that the classifier is a bootstrapping SVM ensemble using the TF weighting schema and count scoring schema while BG_TFIDF_P specifies that it is a bagging SVM ensemble using the TFIDF weighting schema and percentage scoring schema.

5.2 HVSA Indices

While it is possible to use the *s*, *t* and *v* (v_p and v_c) scores individually as an index for segmenting and identifying a HVSA, each method has its own strengths and limitations [27]. It is therefore of interest to analyse if the combination of these scores can generalise the identification task and help to improve the classification result.

5.2.1 The Simple Average Schema

An average value of scores from three methods, namely Fuzzy Match, Twitter LDA, and the best performing SVM ensemble setup, is the first approach for generating a combined HVSA index. The index generated is termed as HVSAave to indicate that it is an average value.

5.2.2 The Linear Regression Schema

A second approach is to adopt a Linear Regression (LR) model to learn about the relationship among scores from the three classifiers (Fuzzy Match, Twitter LDA and SVM ensemble) and its generated ranking position. In other words, we are assessing if it is feasible to use a LR model to predict the rank position using the scores of the three classifiers. HVSAreg indicates that the HVSA index is derived through LR and this value is a predicted ranking position that is different compared to other scores or the HVSAave index, as the smaller the HVSAreg, the better a rank is given to a follower. It is calculated as follows:

$$HVSAreg_f = \beta_0 + \beta_1(s\,score)_f + \beta_2(t\,score)_f + \beta_3(v\,score)_f + \varepsilon_{HVSAreg_f} \quad (5)$$

where,

f = followers

$HSVAreg_f$ = HVSA index generated by LR for follower f

$(s\,score)_f$ = score from FM for follower f

$(t\,score)_f$ = score from TLDA for follower f

$(v\,score)_f$ = score from the best SVM ensemble for follower f

$\varepsilon_{HSVAreg_f}$ = Residual error terms

6. Experimental Setup and Evaluation

This section presents the various setups and metrics used for performance evaluation, ranging from traditional performance metrics such as precision, recall and F measure to evaluation methods adopted from IR research such as Precision@k, Average Precision@k and Average Precision@all. A pooling strategy for extracting unseen testing data is also described.

6.1 Performance Metrics

The typical accuracy metric in statistical analysis of a binary classification, which takes into consideration the true positive (TP) and true negative (TN), has known issues in terms of reflecting the performance of a classifier [31]. Therefore, we have used precision, recall and F measure as performance metrics in this work.

The equations of precision, recall and F measure are as follows:

$$\textit{precision} = TP/(TP + FP) \tag{6}$$

$$\textit{recall} \text{ or } \textit{True Positive Rate (TPR)} = TP/\left(TP + FN\right) \tag{7}$$

$$\textit{True Negative Rate (TNR)} = TN/\left(FP + TN\right) \tag{8}$$

$$\textit{F measure} = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

where TP, TN, FP and FN represent the true positive, true negative, false positive and false negative, respectively.

In addition, ROC analysis and the associated AUC are also presented, as they are both good indicators to assess the overall classification performance.

6.2 Ranking Evaluation

6.2.1 Precision@k and Average Precision@k

While the various scores discussed in Sections 4 and 5 can be used to rank the list of HVSA members extracted, there is a need to evaluate the performance of each classifier and its corresponding scoring schema. For this reason, additional ranking evaluation methods from IR research have been adopted in this study. Specifically, IR uses Precision@k (P@k) and Average Precision@k (AP@k) to measure the quality of top-k retrieved documents in a query (e.g., a result from a search engine). AP@k offers more insights to the quality as it also handles the sensitivity of ranking besides measuring the number of relevant documents retrieved at the top-k cut-off point [32].

We used the scores mentioned in Figure 2 to rank each follower from the AF testing dataset and a positively annotated follower is considered as a relevant retrieval. Four values of *k* were chosen to assess the quality of the retrieval: 10, 25, 50 and 100. These values represent the top 10, 25, 50 and 100 followers selected by each of the classifiers.

P@k can be calculated using the following equation:

$$P@k = \sum n_c / k \quad (10)$$

where n_c is the number of correctly or positively annotated followers within the top k.

In order to better reflect the ranking of HVSA, an adapted AP@k has been used in this study:

$$AP@k = \sum P@k / k \quad (11)$$

Here, k is used in the denominator to reflect how well a method can retrieve the relevant follower and not merely the number of followers retrieved as covered by P@k. The position of the relevant follower is important especially in a situation when a company has a tight budget to work with and they can only allocate a certain amount of fund for a small k. AP@k will be able to address this more accurately so that the company can maximise the budget available.

Furthermore, we also introduced AP@all to indicate the ability of a classifier in identifying all the relevant followers or positively annotated followers from the AF testing dataset. The denominator of AP@all is the number of manually, positively annotated followers. In the case of samsungsg, it is 63 out of the 300 randomly selected followers. For ilovedealssg, it is 23 out of 100 and for beaquafitness, it is 15 out of 50.

6.2.2 A Pooling Strategy

While most classification studies would end with a performance analysis of the annotated testing data, we extended our assessment of the various classifiers to the whole unannotated followers testing data in order to see if it is feasible to adapt the proposed approach to a real-world application. In other words, in addition to evaluating P@k and AP@k on the AF testing dataset, we also applied them to the UAF testing dataset.

However, as the collection of UAF testing datasets is large, a pooling strategy [33] is needed to measure the relative performance of the various scoring schemas and classifiers. The following steps describe the pseudo process of generating a pooled testing dataset for assessment purposes:

i) Choose a diverse set of ranking or scoring classifiers
ii) Run each classifier to return the top-k followers
iii) Combine all the top-k sets to form a pool for human assessors to judge

For this series of analysis, we set the value of k to 10 with a total of ten classifiers for the pooling strategy: FM, TLDA, BT_TF_P, BT_TFIDF_P, BG_TF_P, BG_TFIDF_P, BT_TF_C, BT_TFIDF_C, BF_TF_C and BF_TFIDF_C. FM denotes Fuzzy Match; TLDA denotes Twitter LDA; BG and BT are the bagging and bootstrapping SVM ensembles, respectively; TF and TFIDF are the weight schemas used; P denotes the percentage scoring schema while C denotes the count scoring schema. BG_TF_P refers to the bagging SVM ensemble using the TF weighting schema and percentage scoring schema, while BT_TFIDF_C refers to the bootstrapping SVM ensemble using the TFIDF weighting schema and count scoring schema. The HVSA indices (HVSAave and HVSAreg) were not used to extract the pooled testing dataset so that we could better measure the quality of derived indices. Detailed processes of the pooling strategy are shown in Figure 5.

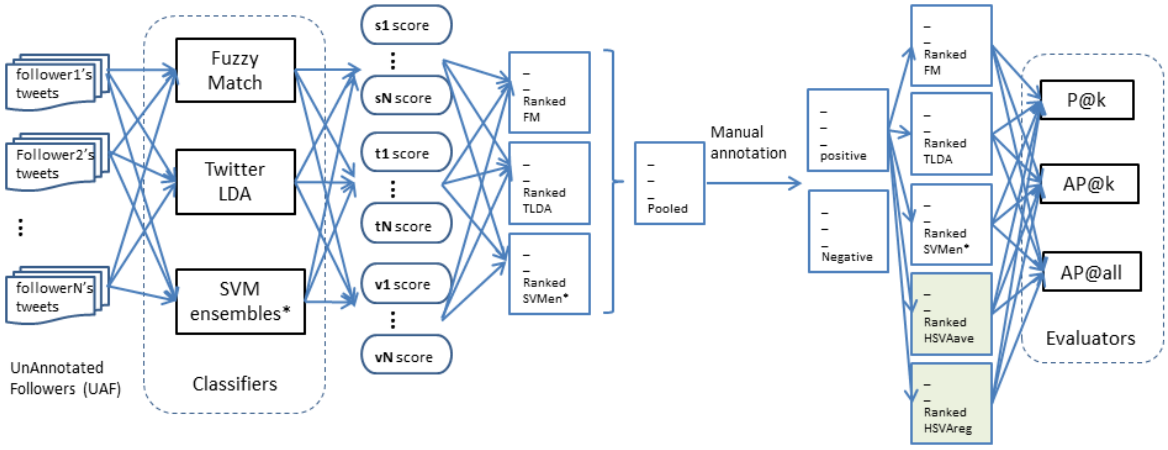

Figure 5. Pooling strategy processes

7. Results

In this section, we present the results obtained for different setups and datasets. We first present results on the SVM ensembles, which include the threshold derived for the count scoring schema as well as results of each variant using both the TF and TFIDF weighting schemas based on training datasets. We then present results in the form of ROC curves and AUC values for the ten classifiers discussed in Section 6.2.2 based on the three AF testing datasets. After that, the performance of each method (including the two HVSA indices) is compared based on IR evaluation metrics (P@k, AP@k, AP@all) and ranking results.

7.1 Thresholds derived for SVM Ensembles with the Count Scoring Schema

A threshold was derived for each of the SVM ensembles considered in this study using the algorithm shown in Figure 4. Table 3 presents the exact threshold value for each ensemble. The corresponding ROC and sensitivity/specificity curve can be found in Figure 6. As observed, datasets with higher AUC values have a higher Youden's index. In general, the TF weighting schema generates better AUC values as compared to TFIDF for all three datasets and different types of ensembles.

Table 3. Threshold and AUC values for various SVM ensembles with the count scoring schema

	samsungsg		**ilovedealssg**		**beaquafitness**	
Method	**Threshold probability value** (Youden's index, accuracy)	**AUC**	**Threshold probability value** (Youden's index, accuracy)	**AUC**	**Threshold probability value** (Youden's index, accuracy)	**AUC**
BG_TF_Count	0.733 (0.621, 0.819)	0.865	0.688 (0.197, 0.649)	0.610	0.767 (0.481, 0.617)	0.783
BT_TF_Count	0.757 (0.612, 0.817)	0.861	0.672 (0.180, 0.625)	0.596	0.771 (0.519, 0.658)	0.774
BG_TFIDF_Count	0.741 (0.559, 0.791)	0.832	0.687 (0.185, 0.644)	0.606	0.758 (0.475, 0.674)	0.762
BT_TFIDF_Count	0.761 (0.549, 0.788)	0.832	0.697 (0.181, 0.641)	0.595	0.765 (0.551, 0.679)	0.760

7.2 Performance of SVM Ensembles on Training Datasets

Next, we evaluated the SVM ensembles using 10 fold cross-validation. The average performance of each setting over 10 runs was recorded and shuffled sampling was used to create samples for each of the settings. Tables 4 and 5 show the results of 10 fold cross-validation for the SVM ensembles using TF and TFIDF as weighting schemas. As can be seen from the tables, slightly higher F measure values are found in the samsungsg dataset for both types of SVM ensembles regardless of the weighting schemas used. This is followed by beaquafitness and ilovedealssg. While the results indicate that using the TFIDF weighting schema has marginally better performance, the difference is minimal for all datasets.

Table 4. Results of 10 fold cross-validation for various SVM ensembles using the TF weighting schema

Dataset	**SVM ensembles**	**Recall**	**Precision**	**F measure**
samsungsg	SVM with bootstrapping sampling	1	0.976	0.988
	SVM with bagging	1	0.977	0.988
ilovedealssg	SVM with bootstrapping sampling	0.976	0.967	0.972
	SVM with bagging	0.970	0.965	0.971
beaquafitness	SVM with bootstrapping sampling	0.971	0.988	0.980
	SVM with bagging	0.971	0.989	0.980

Table 5. Results of 10 fold cross-validation for various SVM ensembles using the TFIDF weighting schema

Dataset	SVM ensembles	Recall	Precision	F measure
samsungsg	SVM with bootstrapping sampling	1	0.980	0.990
	SVM with bagging	1	0.979	0.989
ilovedealssg	SVM with bootstrapping sampling	0.980	0.968	0.974
	SVM with bagging	0.980	0.966	0.973
beaquafitness	SVM with bootstrapping sampling	0.971	0.986	0.979
	SVM with bagging	0.971	0.986	0.979

7.3 Performance Evaluation on the AF Testing Dataset

The ten classifiers as discussed in Section 6.2.2 were assessed using the AF testing dataset. The ROC curves and AUC values can be found in Figures 7 and 8.

As shown in Figure 7a, most classifiers have similar results except FM, which has performed exceptionally well on the samsungsg AF testing dataset. In Figure 7b, the ROC curves for ilovedealssg are slightly different with the percentage scoring schema performing better than the rest. This observation can also be seen in Figure 7c with beaquafitness, where the percentage scoring schema again has performed better than the count scoring schema. In Figure 7c, TLDA is the best performer followed by FM for the beaquafitness AF testing dataset.

Given that some of the ROC curves are very close to each other, it is hard to distinguish which is the best performing SVM ensemble for constructing the HVSA index. We thus calculated the AUC. It is clear from Figure 8 that BG_TF_C performs the best for samsungsg while BG_TF_P and BG_TFIDF_P have the highest AUC values for ilovedealsg and beaquafitness datasets, respectively.

7.4 Ranking Results based on the AF Testing Dataset

As the main purpose of this study is to assess if the scoring schema derived from the various classifiers is capable of ranking the list of HVSA members identified, we also evaluated the results from the AF testing dataset using P@k, AP@k and AP@all discussed in Section 6.2.1. The analyses are presented in two visualisation formats given the difference in datasets and classifiers. Figure 9 shows the comparison between methods based on each of the metrics across different datasets, while Figures 10, 11 and 12 focus on ranking performance of the various methods for each of the datasets.

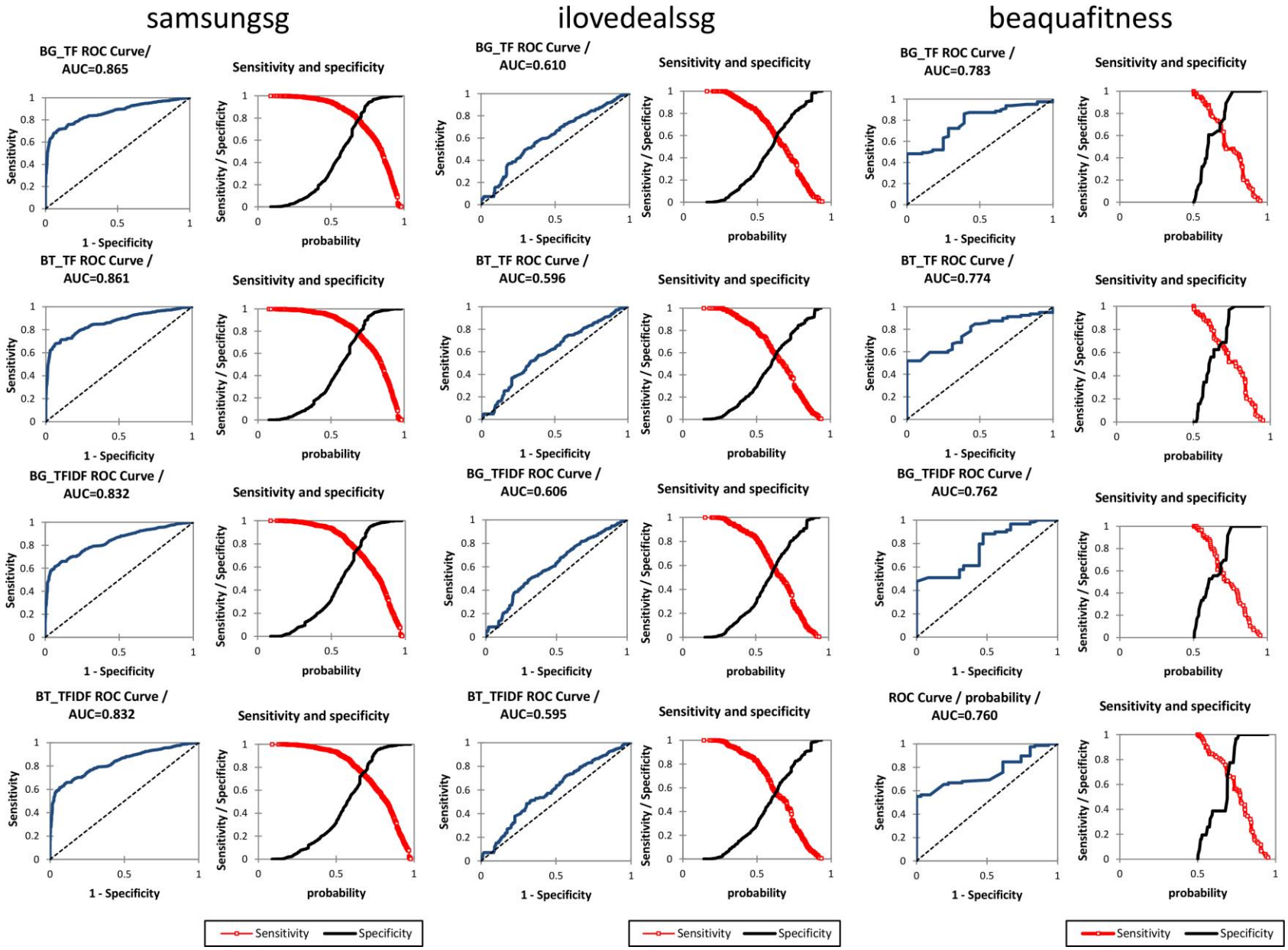
samsungsg
ilovedealssg
beaquafitness
BG_TF ROC Curve/ AUC=0.865
Sensitivity and specificity
BG_TF ROC Curve / AUC=0.610
Sensitivity and specificity
BG_TF ROC Curve / AUC=0.783
Sensitivity and specificity
BT_TF ROC Curve / AUC=0.861
Sensitivity and specificity
BT_TF ROC Curve / AUC=0.596
Sensitivity and specificity
BT_TF ROC Curve / AUC=0.774
Sensitivity and specificity
BG_TFIDF ROC Curve / AUC=0.832
Sensitivity and specificity
BG_TFIDF ROC Curve / AUC=0.606
Sensitivity and specificity
BG_TFIDF ROC Curve / AUC=0.762
Sensitivity and specificity
BT_TFIDF ROC Curve / AUC=0.832
Sensitivity and specificity
BT_TFIDF ROC Curve / AUC=0.595
Sensitivity and specificity
ROC Curve / probability / AUC=0.760
Sensitivity and specificity
Sensitivity
1 - Specificity
Sensitivity / Specificity
probability
Sensitivity
Specificity

Figure 6. ROC curves and sensitivity/specificity plots for the count scoring schema. BG is the bagging SVM ensemble while BT is the bootstrapping SVM ensemble. TF and TFIDF are the weighting schemas used.

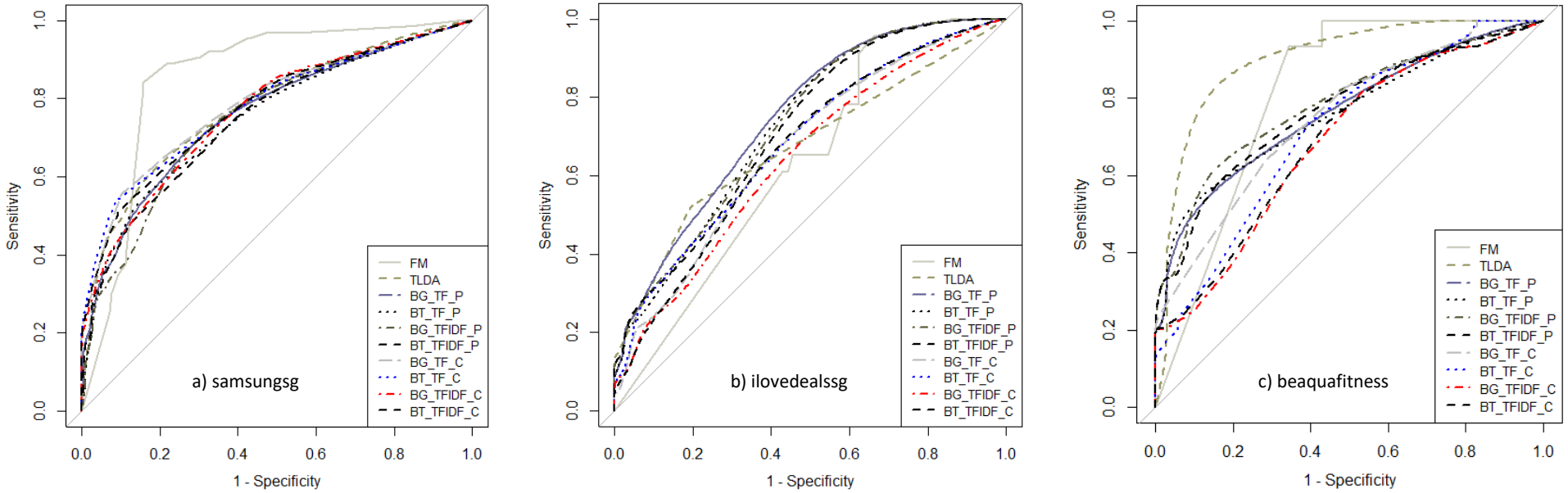

Figure 7. ROC curves of various classifiers on a) samsungsg, b) ilovedealssg, and c) beaquafitness AF testing datasets

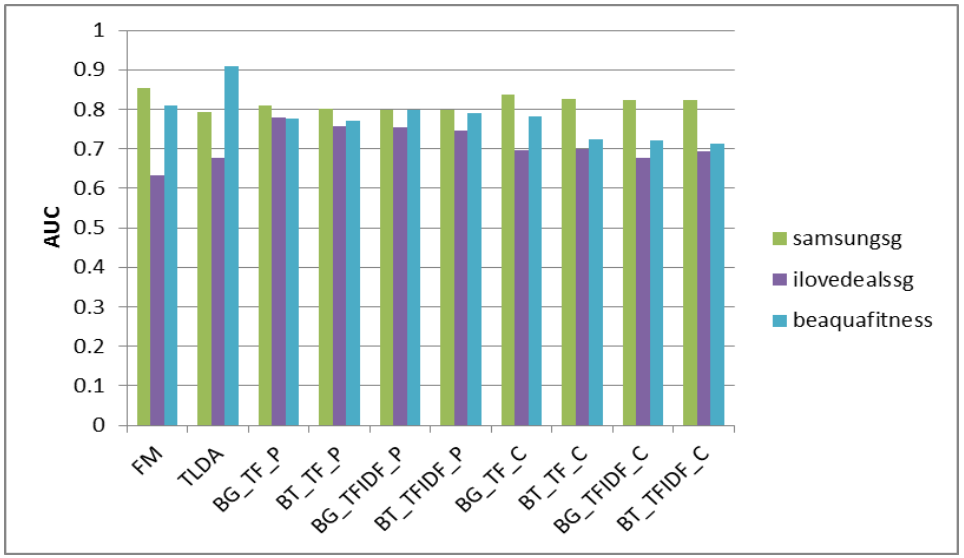

Figure 8. AUC values of various classifiers for samsungsg, ilovedealssg and beaquafitness

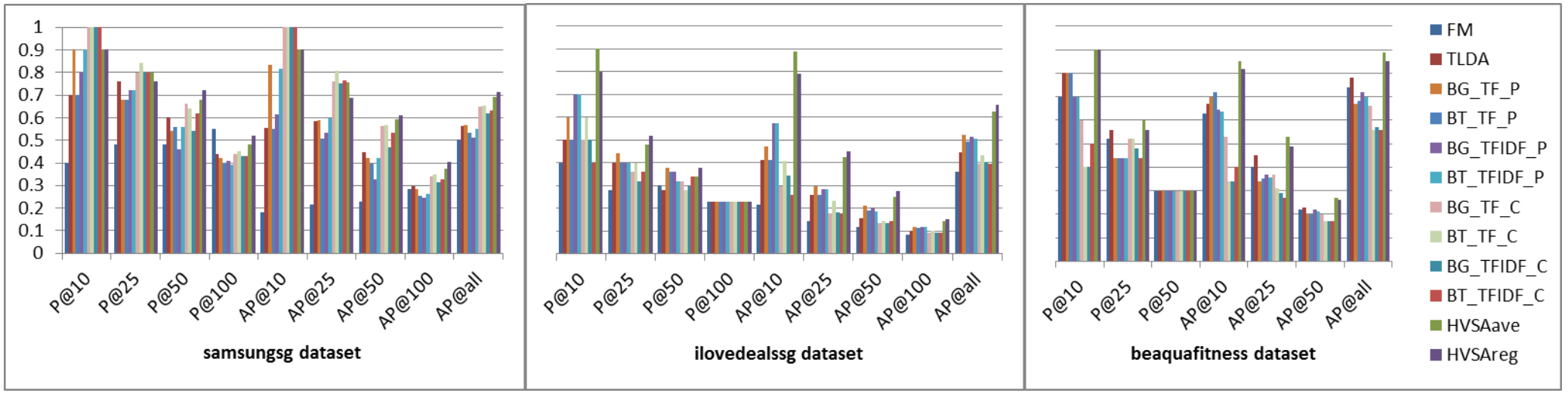

Figure 9. Evaluation results based on the metrics (P@k, AP@k, AP@all) using different AF testing datasets

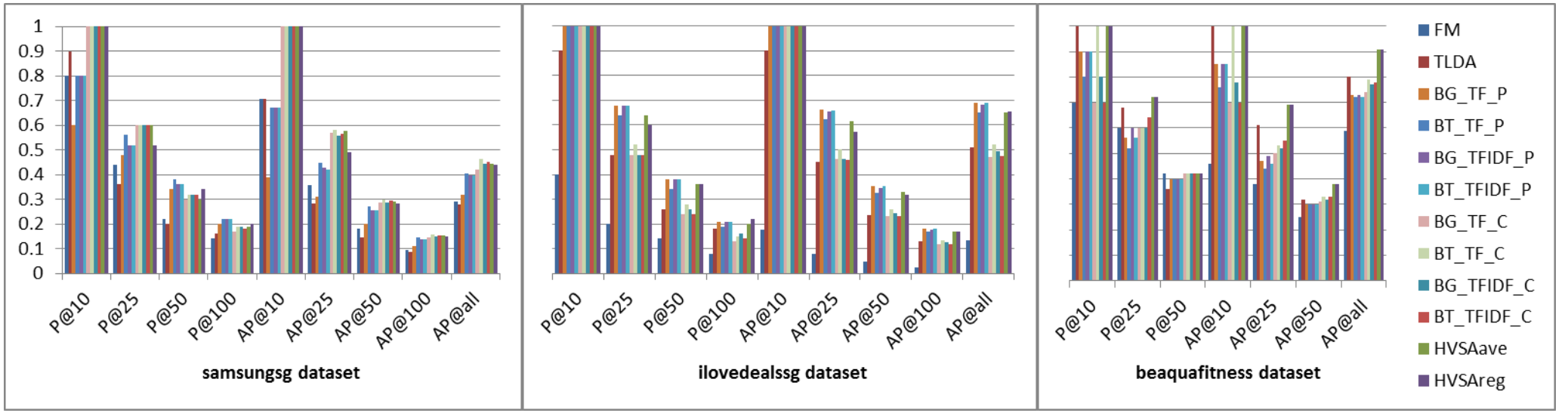

Figure 13. Evaluation results based on the metrics (P@k, AP@k, AP@all) using the pooling strategy on different datasets

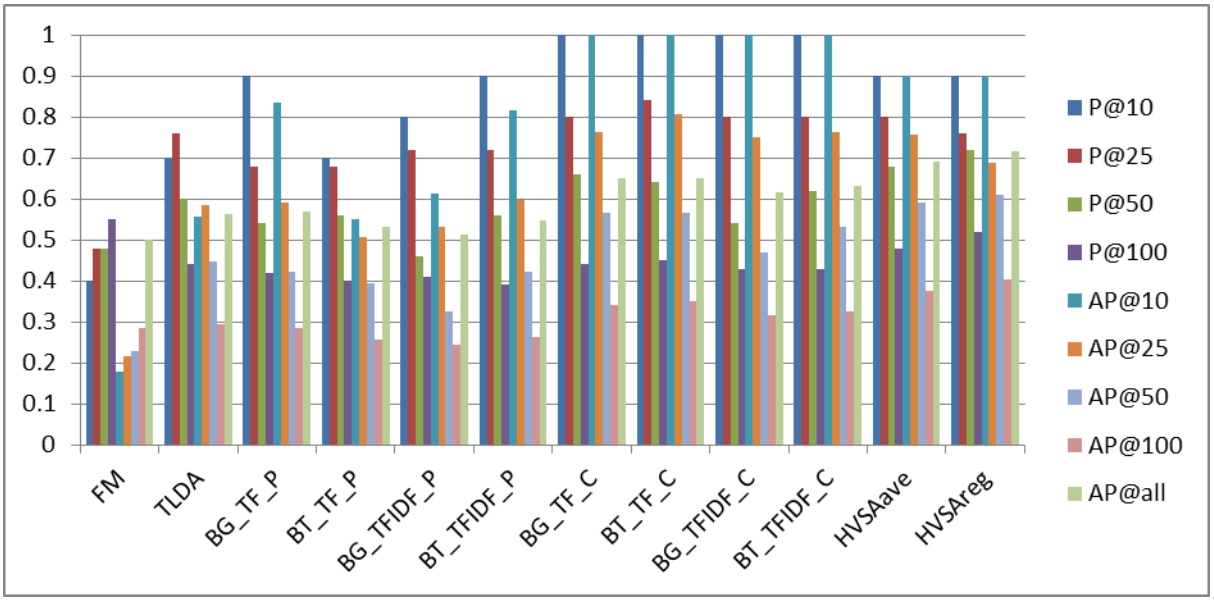

Figure 10. Ranking performance of various methods based on the samsungsg AF testing dataset

In general, we see in Figure 9 that samsungsg has higher scores compared to ilovedealssg and beaquafitness. The lower scores of ilovedealssg and beaquafitness are mainly due to the heterogeneity and diversity of their tweets. For example, the lexical diversity of words in the processed tweets of ilovedealssg is 0.535 followed by beaquafitness at 0.445, while samsungsg's is only 0.365. These values were calculated by taking unique tokens (i.e., words) of the text divided by the total number of tokens [34]. The heterogeneity of a dataset may pose great challenges for any classifier as it is more difficult to find a representative training dataset for accurate classification.

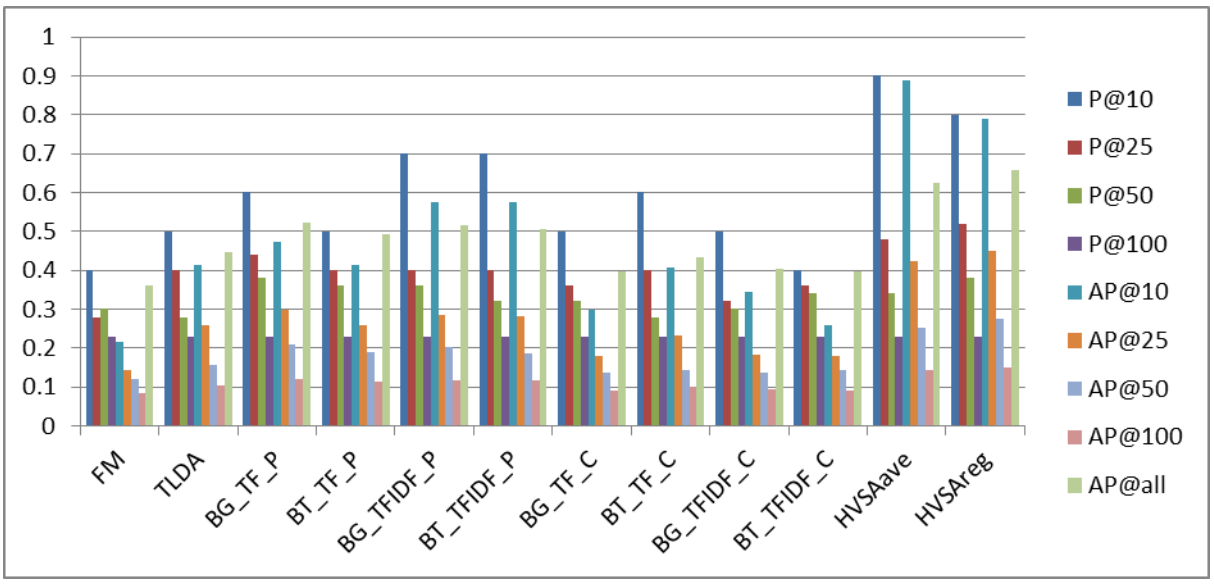

Figure 11. Ranking performance of various methods based on the ilovedealssg AF testing dataset

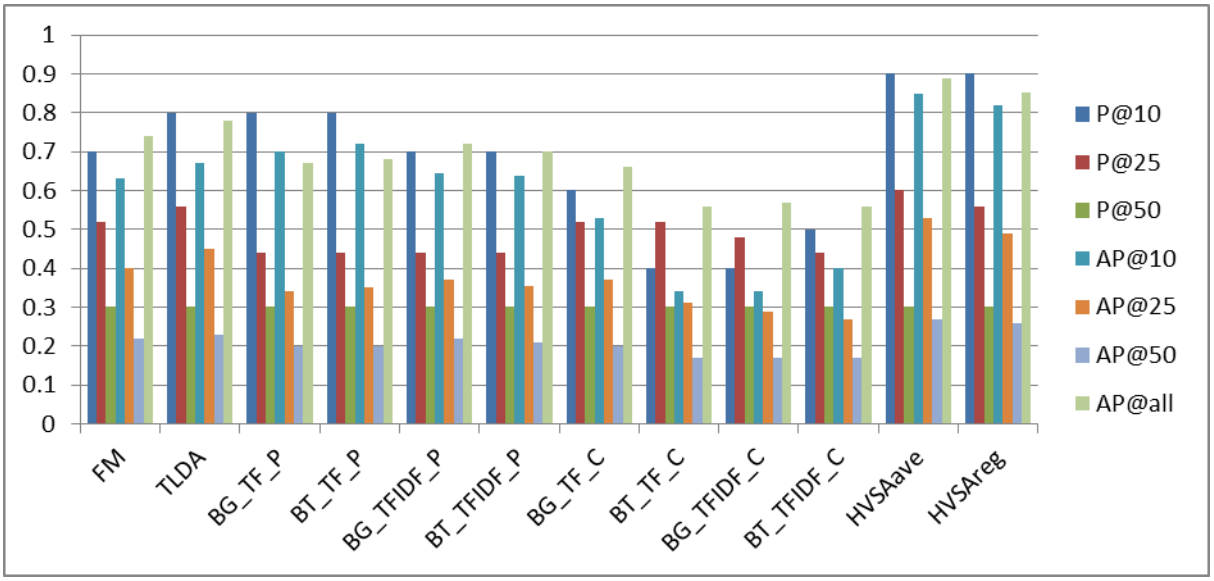

Figure 12. Ranking performance of various methods based on the beaquafitness AF testing dataset

From Figure 9, we also see that AP@k is indeed a more sensitive evaluator than P@k. As the AF testing dataset of ilovedealsg has 100 followers, P@100 shows the same precision value for all methods, indicating that all the relevant followers have been retrieved for the whole dataset. AP@100, on the other hand, is sensitive enough to show the differences. Both the HVSA indices have performed better compared to others, implying that these HVSA indices are more capable of identifying the correct ranking of followers. Similar observations can be said of beaquafitness for P@50 and AP@50. P@100 and AP@100 are omitted for beaquafitness due to the smaller size of its dataset.

As a whole, the HVSA indices have higher scores for AP@all on all three datasets. This suggests that the derived indices are leveraging on the combined strengths of the different classifiers, namely FM, TLDA and the SVM ensemble. While it is clearly shown in Figures 11 and 12 that the HVSA indices are preferred in the ilovedealssg and beaquafitness datasets, SVM ensembles with the count scoring schema are performing better in the samsungsg dataset as shown in Figure 10.

From Figures 10, 11 and 12, we observe that the count scoring schema has better results in the samsungsg dataset while the percentage scoring schema has achieved higher values in the ilovedealssg and beaquafitness datasets. There is no striking difference for weighting schemas of TF or TFIDF.

7.5 Results from the Pooling Strategy

7.5.1 Ranking Results using the Pooling Strategy

Figure 13 shows the results of the three datasets using the pooling strategy. It is obvious that the HVSA indices perform reliably well as compared to other methods. This indicates that the indices are capable of identifying HVSA regardless of dataset types.

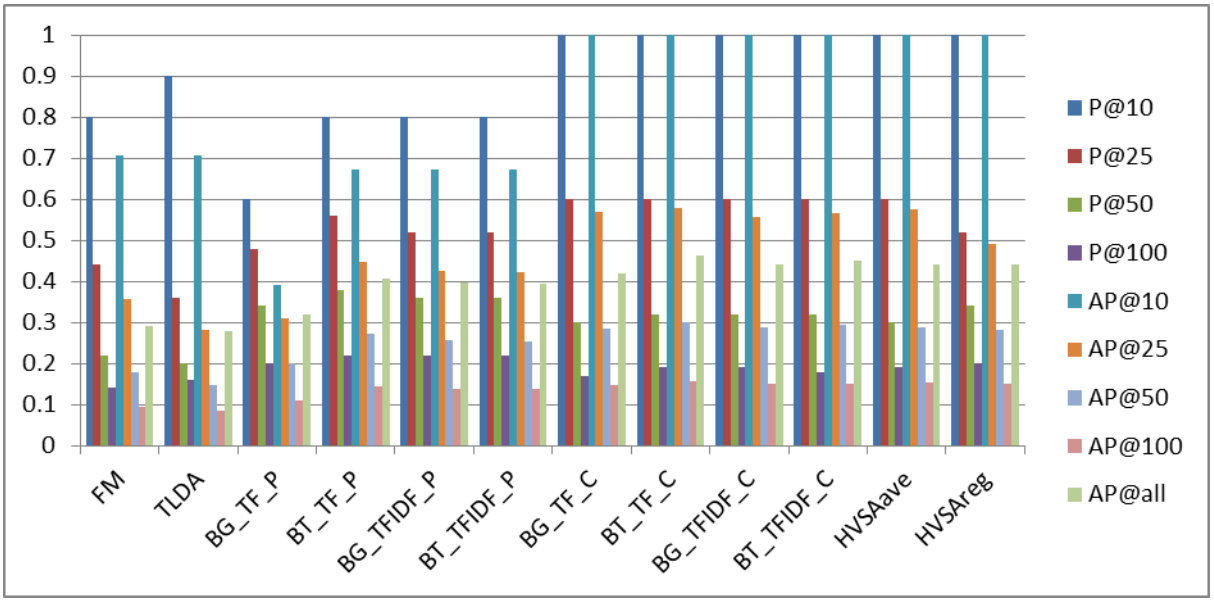

Figure 14. Ranking performance of various methods based on samsungsg using the pooling strategy

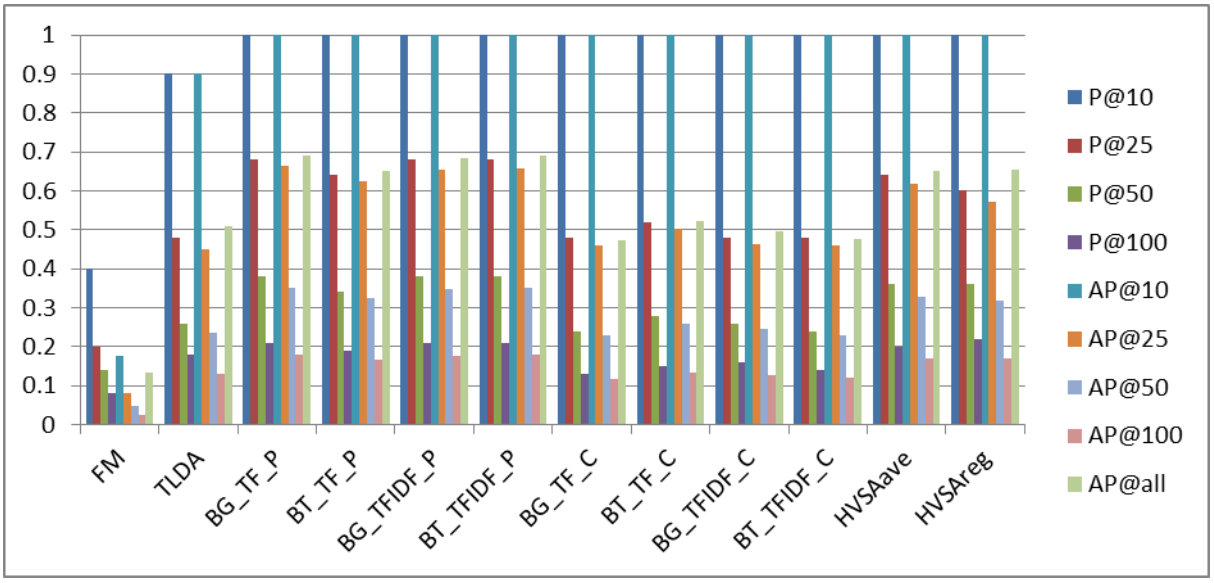

Figure 15. Ranking performance of various methods based on ilovedealssg using the pooling strategy

From Figures 14, 15 and 16, we can see that the ranking results with the pooling strategy are consistent with the ranking results using the AF testing dataset, in which the percentage scoring schema is preferred in the ilovedealssg and beaquafitness datasets while the count scoring schema is the choice in the samsungsg dataset. However, the difference in the scoring range is not as big compared to that of the AF

testing dataset. In other words, the pooling strategy may minimise the effect caused by the different nature of datasets, and concentrate on evaluating the efficacy of the various methods (as it is based on the whole unannotated dataset).

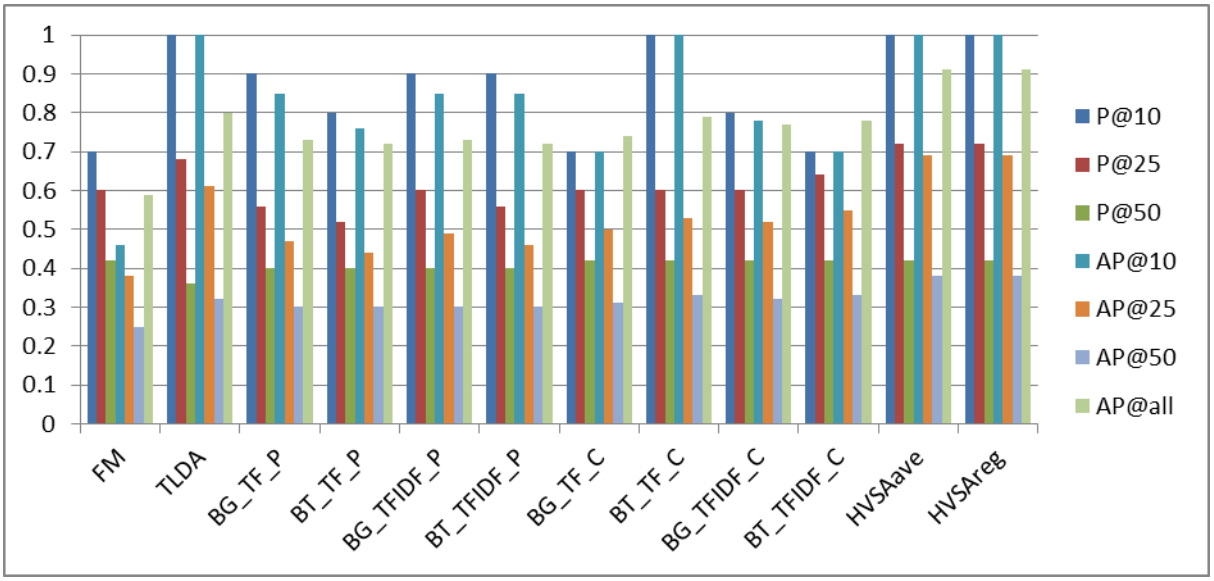

Figure 16. Ranking performance of various methods based on beaquafitness using the pooling strategy

7.5.2 Audience Segmentation through Twitter LDA

While having a ranked list of followers will help in engaging the potential social audience rather than spamming every one of them, it would be even more beneficial if the engagement of top-k audience members (i.e., the HVSA) can be personalised so that appropriate offers can be sent to the right person. Twitter LDA was therefore used to analyse the followers identified from the pooling strategy as it provides the ability to segment the top-k audience members according to their domains of interest. A 10-topic model was applied and the list of topics is shown in Table 6.

Some of the topics have been omitted, as they are about daily mundane comments. The scope of interests from ilovedealssg followers can be vast, as their interests include spa, beauty regime, property, travel and hotel, services and products, and holiday resort. While much of the content from samsungsg followers is related to its products, they have varying interests ranging from phones, cameras, televisions to laptops. Besides that, topic modelling also reveals that samsungsg followers do read their Samsung Village blog and it may be rewarding to engage their followers through information or promotion on Korean's entertainment people like shine or jonghyun. The results presented here are consistent with

Table 1 as the Entertainment_Culture topic is also found in samsungsg's content. Most of the followers of beaquafitness found from the pooling strategy are active in exercising and paying attention to healthy living and fitness regimes. Some specialised terms such as aquacycling and marathon have been identified among the followers and the distinctive grouping shows that its followers are passionate about these topics, which may explain why TLDA performs best in this dataset (see Figures 7c and 8).

Table 6. Topic modelling based on the pooling strategy results

Followers from the pooling strategy	**Topic ID**	**Top words in each Topic**
ilovedealssg	0	resort, spa, city, grand
	2	perm, lash, threading, brow, women
	3	deals, dubai, apartment
	6	japan, korea, Taiwan, hotel
	7	worth, whitening, teeth, toothbrush, bag, facial, massage
	9	resort, batam, beach, nongsa, turi, island
samsungsg	0	galaxy, android, samsung, note
	1	iphone, ipad, htc
	3	camera, casio, nokia, smart
	4	shinee, omg, jonghyun, media
	6	samsung, village, blog, electronics
	7	samsung, tv, led, phablet, lcd
	8	phone mobile, palm, motorola, treo, laptop
beaquafitness	0	pool, legs, arms, exercise, aqua
	1	aqua, aquacycling, cycling, classes, aquaspin
	2	water, body, foods, fitness, weight, training
	3	coaching, swim, freestyle, training
	5	calories, fat, diet, eat, intake
	7	sleep, life, healthyliving, motivation
	8	running, marathon, trailrunning

Table 7. Audience segmentation using TLDA and the HVSA index on samsungsg followers

Topic ID	3	4	7	8
Followers* (where the number is the ranking by HVSAave)	Follower6 Follower14	Follower1 Follower3 Follower4 Follower9 Follower10 Follower12 Follower17	Follower2 Follower3 Follower7 Follower9 Follower10 Follower13	Follower2 Follower5 Follower7 Follower8 Follower9 Follower10 Follower12

*Analysis is only done on the top 25 followers of HVSAave

With results from the audience segmentation based on TLDA, we are now able to identify the top social audience members for each relevant individual topic – see Table 7. It is not surprising to find a follower being identified in more than one topic (for example, Follower9 and Follower10 are both found

in Topics 4, 7 and 8) and thus, by combining topics and ranking of the HVSA, a more detailed selection is available to aid in better decision-making.

8. Discussion

Previous studies along this line of research (e.g., [11] [24] [27] [35]) have focused on a narrow domain (samsungsg) by exploring some text mining and machine learning methods [11] [24] [35] cum introducing a simple average schema for ranking purposes without considering the optimal cut-off threshold [27]. Extensive work, which includes the use of more datasets of different nature and domains (ilovedealssg and beaquafitness), an attempt to find the optimal cut-off point using Youden's index, the adapted IR ranking algorithm, introduction of the regression HVSA schema and comprehensive evaluation methods, has been carried out in this paper to investigate if the HVSA index can indeed be used to identify a target audience. The findings are largely positive, and have provided significant insight into the potential value of this research for real-world targeted marketing. In this section, we discuss the classification results first based on the characteristics of the datasets, then the scoring schemas, and after that the methods ranging from SVM ensembles to FM and TLDA. Finally, we wrap up the section with some remarks about the IR evaluation metrics and HVSA indices.

The SVM ensembles have achieved higher AUC values compared to FM or TLDA on ilovedealssg (as shown in Figure 8). This is partly due to the fact that SVM ensembles are leveraging the diversity of the dataset for their advantage. However, other research [36] has also shown that too much diversity may impact on the accuracy, and hence it is a delicate task to achieve a balance in both. A further analysis on the annotation of the three AF testing datasets showed that there were 74% of the followers in ilovedealssg sharing about mundane comments while only 37% and 26% were doing so in samsungsg and beaquafitness, respectively. In contrast, 33% of samsungsg followers and 30% of beaquafitness followers shared content related to their owners, whereas only 23% of ilovedealssg followers did so. Although such diversity was considered in the construction of the training dataset for ilovedealssg, the heterogeneity of the dataset may not be captured well in the semi-supervised learning process due to the wide ranging domains of daily mundane sharing and hence it is inevitable that most classifiers have lower scores on the ilovedealssg dataset compared to those for samsungsg and beaquafitness. Furthermore, with samsungsg being a mobile technology company where the domain and vocabulary used are well-defined, it is

understandable that most classifiers, including the FM method, are able to differentiate the classes better. Similar results have been observed for the beaquafitness dataset (as shown in Figure 8) with FM and TLDA achieving better AUC values than others. This is most likely due to the specialised and non-ambiguous terms (describing the various equipment and activities) used in the beaquafitness dataset, although its lexical diversity is higher than samsungsg (see Section 7.4).

It is interesting to observe that the count scoring schema has consistently performed well in the samsungsg dataset (see Figures 10 and 14), while the percentage scoring schema stands out from the rest for the ilovedealssg (see Figures 11 and 15) and beaquafitness (see Figures 12 and 16) datasets. This may likely be due to the fact that the count scoring schema is more sensitive in identifying followers who share similar content, as the total number of tweets shared is not considered. Being an account with a more defined domain, samsungsg can benefit from this approach. However, this also indicates that it can be challenging to identify followers who may have shared very little on the content of interest for ilovedealssg, as most of the positively identified followers are those who have shared more similar content. In fact, a detailed analysis on the top HVSA members identified from the pooling strategy showed that many followers of ilovedealssg are Twitter users or business owners who share on various deals and promotions such as I_LOVE_Discount and DEALGuruSG.

The bagging SVM ensemble was the best performer in [24]. However, it has not shown any significant advantage in ranking evaluation in both the AF testing and unseen datasets from the pooled strategy for identifying top-k followers from samsungsg, ilovedealssg and beaquafitness. The results from this study show that both bootstrapping and bagging SVM ensembles have similar performances and the distinctive difference lies in the scoring schema used to represent followers. Besides that, additional experiments with ten single SVMs constructed using ten randomly selected subsets of the training datasets showed significant differences in classification accuracies among the SVMs. The suggestion is therefore to use an ensemble approach to handle imbalance in the training datasets as it is able to minimise bias or over-representation of any class.

While FM has achieved one of the largest AUCs based on the samsungsg dataset (see Figure 8), it is not able to identify the top ranking HVSA members with precision (see Figure 10). It also does not perform well in the ilovedealssg dataset, mostly due to the inability of fuzzy keyword match on words

from diverse domains. In other words, FM would work well for account owners involving in businesses dealing with specific products, e.g., samsungsg and beaquafitness, however, it is not designed to handle matching of synonyms such as “resort” and “spa” found in ilovedealssg.

Interestingly, TLDA has outperformed others in the beaquafitness dataset. This is mainly due to the fact that, although the lexical diversity of content shared in beaquafitness is higher than samsungsg, the terms used are more specialised and specific to the domain compared to ilovedealssg, and thus topic modelling can be done more successfully on the dataset. It is worthwhile highlighting that, similar to FM in the samsungsg dataset, TLDA does not perform as well as HVSA indices in the beaquafitness dataset (see Figure 12). It is hence advisable to adopt evaluation metrics such as AP@k for identifying top ranking followers and rely on other scoring schemas such as HVSA indices for better identification. Both HVSAave and HVSAreg indices perform relatively well and stable across the three datasets, with AP@all achieving the best performance (see Figure 9), and hence it is suggested that a combined HVSA index (constructed either through simple average or LR), which leverages on different approaches (e.g., FM, TLDA and SVM ensembles), is recommended for ranking the HVSA.

Even though we have only conducted investigation using k=10, 25, 50 and 100, the results have shown that IR evaluation metrics can be applied in ranking of HVSA with satisfactory outcomes instead of relying on classic performance metrics such as F measure or the AUC, which may not be able to identify the best classifiers for the task. Besides that, it is interesting to highlight that HVSA indices derived in this study can outperform other methods regardless of the type of datasets. In addition, we have also investigated the use of audience segmentation through TLDA and we believe that by ranking the HVSA with different segments, it is more beneficial as a company is then able to narrow down the group of HVSA members that match its latest marketing plan instead of choosing randomly from the vast number of followers.

9. Conclusion

In this paper, we have demonstrated that it is possible to identify the top-k followers to aid a company in making decisions when doing business on social media. We have also shown that the HVSA indices derived are capable of retrieving HVSA members with high precision in a range of datasets. In addition, with a combination of semi-supervised (Twitter LDA) and supervised (SVM ensembles) learning

approaches, we have developed a mechanism that is able to identify the HVSA from a list of followers with minimal annotation effort, rank the list of HVSA members and segment them according to their interests so that the company can devise different engagement or promotion plans to target different groups of audience members more effectively.

References

[1] '2013 Fortune 500 - UMass Dartmouth'. [Online]. Available: http://www.umassd.edu/cmr/socialmediaresearch/2013fortune500/. [Accessed: 29-Jul-2014].

[2] M. B. Goodman, N. Booth, and J. A. Matic, 'Mapping and leveraging influencers in social media to shape corporate brand perceptions', *Corp. Commun. Int. J.*, vol. 16, no. 3, pp. 184–191, 2011.

[3] 'Microsoft's Internet Explorer Influencer Campaign Backfires: Another High-Profile Example Of Why Details Matter'. [Online]. Available: http://marketingland.com/microsofts-internet-explorer-influencer-campaign-backfires-another-high-profile-example-details-matter-87891. [Accessed: 16-May-2015].

[4] M. Pennacchiotti and A.-M. Popescu, 'Democrats, republicans and starbucks afficionados: user classification in twitter', in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 430–438.

[5] T. Yang, D. Lee, and S. Yan, 'Steeler nation, 12th man, and boo birds: classifying Twitter user interests using time series', presented at the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 684–691.

[6] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, 'Classifying latent user attributes in twitter', in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 2010, pp. 37–44.

[7] K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino, 'Twitter user profiling based on text and community mining for market analysis', *Knowl.-Based Syst.*, vol. 51, pp. 35–47, 2013.

[8] W. Rao, L. Chen, and I. Bartolini, 'Ranked content advertising in online social networks', *World Wide Web*, pp. 1–19, 2014.

[9] 'Digg - What the Internet is talking about right now'. [Online]. Available: http://digg.com/. [Accessed: 10-Feb-2015].

[10] L. Tang, Z. Ni, H. Xiong, and H. Zhu, 'Locating targets through mention in Twitter', *World Wide Web*, pp. 1–31, 2014.

[11] S. L. Lo, D. Cornforth, and Raymond,Chiong, 'Identifying the High-Value Social Audience from Twitter through Text-Mining Methods', in *Proceeding of Adaptation, Learning and Optimization Series*, Singapore, 2014.

[12] D. Harman, 'Information retrieval evaluation', *Synth. Lect. Inf. Concepts Retr. Serv.*, vol. 3, no. 2, pp. 1–119, 2011.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, 'Comparing twitter and traditional media using topic models', in *Advances in Information Retrieval*, Springer, 2011, pp. 338–349.

[14] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[15] M. Michelson and S. A. Macskassy, 'Discovering users' topics of interest on twitter: a first look', in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, 2010, pp. 73–80.

[16] L. Hong, A. S. Doumith, and B. D. Davison, 'Co-factorization machines: modeling user interests and predicting individual decisions in twitter', in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 557–566.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, 'Latent dirichlet allocation', *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[18] Y. Zhang and M. Pennacchiotti, 'Predicting purchase behaviors from social media', in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1521–1532.
[19] J.-W. van Dam and M. van de Velden, 'Online profiling and clustering of Facebook users', *Decis. Support Syst.*, vol. 70, pp. 60–72, 2015.
[20] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, 'Adapting ranking SVM to document retrieval', presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 186–193.
[21] R. Fagin, A. Lotem, and M. Naor, 'Optimal aggregation algorithms for middleware', *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 614–656, 2003.
[22] I. F. Ilyas, G. Beskales, and M. A. Soliman, 'A survey of top-k query processing techniques in relational database systems', *ACM Comput. Surv. CSUR*, vol. 40, no. 4, p. 11, 2008.
[23] 'Thomson Reuters | Open Calais'. [Online]. Available: http://new.opencalais.com/. [Accessed: 13-May-2015].
[24] S. L. Lo, R. Chiong, and D. Cornforth, 'Using Support Vector Machine Ensembles for Target Audience Classification on Twitter', 2015.
[25] G. Kondrak, D. Marcu, and K. Knight, 'Cognates can improve statistical translation models', presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2, 2003, pp. 46–48.
[26] M.-C. Yang and H.-C. Rim, 'Identifying Interesting Twitter Contents Using Topical Analysis', *Expert Syst. Appl.*, 2014.
[27] S. L. Lo, D. Cornforth, and R. Chiong, 'Use of a High-Value Social Audience Index for Target Audience Identification on Twitter', in *Artificial Life and Computational Intelligence*, Springer, 2015, pp. 323–336.
[28] C. J. Burges, 'A tutorial on support vector machines for pattern recognition', *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.
[29] 'Predictive Analytics, Data Mining, Self-service, Open source - RapidMiner'. [Online]. Available: http://rapidminer.com/. [Accessed: 27-Jun-2014].
[30] W. J. Youden, 'Index for rating diagnostic tests', *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[31] M. Sokolova, N. Japkowicz, and S. Szpakowicz, 'Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation', in *AI 2006: Advances in Artificial Intelligence*, Springer, 2006, pp. 1015–1021.
[32] S. Robertson, 'A new interpretation of average precision', presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 689–690.
[33] J. Zobel, 'How reliable are the results of large-scale information retrieval experiments?', presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 307–314.
[34] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, Inc., 2013.
[35] S. L. Lo, D. Cornforth, and R. Chiong, 'Effects of Training Datasets on both the Extreme Learning Machine and Support Vector Machine for Target Audience Identification on Twitter', presented at the Extreme Learning Machine, Singapore, 2014.
[36] Z. Xiao, P. Li, Y. Fu, and M. Lu, 'Does More Randomness Help Increasing Diversity in Ensemble Learning?', *J. Converg. Inf. Technol.*, vol. 7, no. 19, 2012.