2-2015

Use of a high-value social audience index for target audience identification on Twitter

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

David CORNFORTH

Raymond. CHIONG

Citation

1

Use of a High-Value Social Audience Index for Target Audience Identification on Twitter

Siaw Ling Lo, David Cornforth, and Raymond Chiong

School of Design, Communication and Information Technology
Faculty of Science and Information Technology
The University of Newcastle, Callaghan, NSW 2308, Australia
siawling.lo@uon.edu.au,
{david.cornforth,raymond.chiong}@newcastle.edu.au

Abstract. With the large and growing user base of social media, it is not an easy feat to identify potential customers for business. This is mainly due to the challenge of extracting commercially viable contents from the vast amount of free-form conversations. In this paper, we analyse the Twitter content of an account owner and its list of followers through various text mining methods and segment the list of followers via an index. We have termed this index as the High-Value Social Audience (HVSA) index. This HVSA index enables a company or organisation to devise their marketing and engagement plan according to available resources, so that a high-value social audience can potentially be transformed to customers, and hence improve the return on investment.

Keywords: Twitter, topic modelling, machine learning, audience segmentation.

1 Introduction

Twitter and Facebook are no longer a fad but a gold mine for any business to reach out to potential customers, since both platforms have a huge active user base of over 1.28 billion [1]. More companies are putting emphasis upon, or have started, their social media marketing strategy plan in hopes of standing out from the increasingly crowded social space and attracting prospective customers from the audience. A recent study [2] found that nearly 80% of consumers would more likely be interested in a company due to its brand's presence on social media. It is therefore not surprising that 77% of the Fortune 500 companies have active Twitter accounts and 70% of them maintain an active Facebook account to engage with their potential customers [3].

While a company can rely on incentive referrals to boost its fans' or followers' count, this approach may only provide short-term gain. Furthermore, values from the fans' count and number of retweets are not able to directly provide any actionable insights in customer engagement, although they can be used as one of the measurements in a social media campaign. With the growing "sophistication" of social media users, it can be rewarding for any company if personalised services and quality contents can be offered directly to the fans or followers. Mass marketing can no longer be

S.K. Chalup et al. (Eds.): ACALCI 2015, LNAI 8955, pp. 323–336, 2015.

justified by the effort and amount of money spent. Moreover, there is a thin line between broadcasting a general message and spamming, so instead of attracting a greater audience, there is a high risk of losing current customers. Hence, it makes sense to identify a target audience in order to maximise marketing efficiency and improve the return on investment.

Every profit-oriented business would aim to increase profit, build a long lasting brand name, grow its customer base and further engage its current customers. It is therefore essential to understand the needs and behaviours of customers. This understanding can be achieved through different means and at different levels of detail. Most companies define segmentation of their customers according to their traits or behaviours. All other marketing efforts, such as customer engagement activities, are targeted and measured according to the segmentation.

However, this segmentation is typically restricted to customer relationship management (CRM) or transaction data obtained either through customer surveys or tracking of product purchases, to understand the customer demand. Demographic variables, RFM (recency, frequency, monetary) and LTV (lifetime value) are the most common input variables used in the literature for customer segmentation and clustering [4, 5]. With the rise of social media, there is an emerging model of CRM called "social CRM" [6], which addresses the effect of social media on CRM and its pitfalls. Recent work in this area focuses on a framework of the social CRM model [7], where one of the challenges is violation of customer privacy due to the linking of disparate sources of data. In view of this, it is of interest to analyse alternative approaches (such as using the content shared on social media) in order to determine whether these approaches can be used to complement the traditional CRM in identifying a target or high-value social audience for a company or a product.

There have been efforts in deriving or estimating demographics information [8, 9] from the available social media data, but it may not be feasible to use this set of information directly in targeted marketing, as temporal effects and types of products to be targeted are usually not considered. Besides that, demographic attributes such as age, gender and residence areas on social media platforms may not be updated and hence may result in a misled conclusion. Recently, eBay has expressed that, due to viral campaigns and major social media activities, marketing and advertising strategies are evolving. Targeting specific demographics through segmentation, although still has its value, is being replaced by new strategies. For example, eBay is focusing on "connecting people with the things they need and love, whoever they are" [10]. Other research on predicting purchase behaviour from social media has shown that Facebook categories such as likes, and text analysis methods such as n-grams, significantly outperform demographic features shared on Facebook [11]. Due to the privacy policy of Facebook profiles, our work focuses on Twitter, where most of the contents and activities shared online are open and available.

Considering the vast amount of available social media data, it is not practical to annotate tweets manually to construct a training data for an analysis method. Consequently, the contents from a Twitter account owner are used to extract a list of seed words as the positive training data for the various text mining methods considered in this paper, which include keyword matching through Fuzzy Keyword Match, statistical topic modelling through Latent Dirichlet Allocation (LDA) [12], and machine learning through the Support Vector Machine (SVM) [13]. The scores derived from each of the methods are analysed and combined for the construction of a High-Value Social Audience (HVSA) index, which can subsequently be used in segmenting the list of followers of the account owner for targeted engagement and marketing.

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

- To the best of our knowledge, our work in this paper is the first attempt to define an index capable of identifying a high-value social audience for segmenting the list of followers of a Twitter account owner.
- From the result observation, the content shared by the account owner can be used for customer segmentation as it contains information that is relevant to identify the target audience.

2 Methods

The focus of this research was to establish an index that can be used to segment the social audience (or the list of followers) of a Twitter account owner using the content shared by the account owner. The architecture of our system is given in Fig. 1.

Tweets from various parties - owners, followers, and owners from other domains - were cleaned and pre-processed before preparing for seed words generation and SVM training and testing datasets. The owner's tweets were used as the positive training data while tweets of owners from other domains were extracted as the negative training data. 10-fold cross validation was applied on both the positive and negative training data for the SVM model before the classification of followers' tweets (or the testing data) was conducted. The seed words generated were used by both Fuzzy Keyword Match and Twitter LDA. A string similarity score derived from Dice coefficient was calculated through a fuzzy comparison with the seed words on the testing data. A list of topics was learned from testing data using Twitter LDA and followers with relevant topic numbers were identified. Details of each component are described in the following sections.

2.1 Data Collection

We have used the Twitter Search API [14] for data collection. As the API is constantly evolving with different rate limiting settings, our data gathering has been done through a scheduled program that requests a set of data for a given query. For this particular research, we selected Samsung Singapore or "samsungsg" (its Twitter username) as the subject or brand. At the time of this work, there were 3,727 samsungsg followers. In order to analyse the contents or tweets of the account owner, the last 200 tweets by samsungsg were extracted. The time of tweets was from 2 Nov 2012 to 3 Apr 2013. For each of the followers, the API was used to extract their tweets, giving a total of 187,746 records.

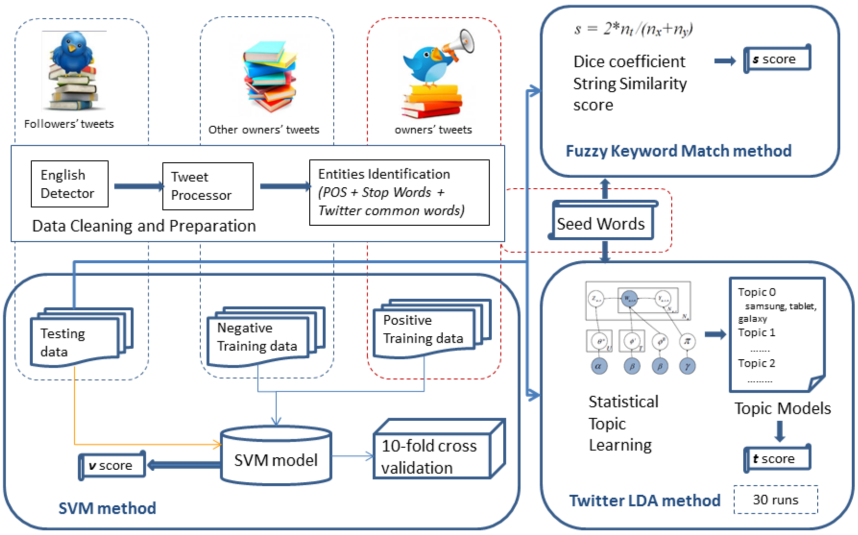

Fig. 1. The system architecture

2.2 Data Cleaning and Preparation

Tweets are known to be noisy and often mixed with linguistic variations. It is hence very important to clean up the tweet content prior to any content extraction:

- Non-English tweets were removed using the Language Detection Library for Java [15];
- URLs, any Twitter's username found in the content (which is in the format of @username) and hashtags (with the # symbol) were removed;
- Each tweet was pre-processed to lower case.

As tweets are usually informal and short (up to 140 characters), abbreviations and misspellings are often part of the content and hence the readily available Named Entity Recognition (NER) package may not be able to extract relevant entities properly. As such, we derived an approach called Entities Identification, which uses Part-of-Speech (POS) [16] tags to differentiate the type of words. In this approach, all the single nouns are identified as possible entities. If the tag of the first fragment detected is 'N' (noun) or 'J' (adjective) and the consecutive word(s) is of the 'N' type, these words will be extracted as phrases. This approach was then complemented by another process using the comprehensive stop words list used by search engines (http://www.webconfs.com/stop-words.php) in addition to a list of English's common words (preposition, conjunction, determiners) as well as Twitter's common words (such as "rt", "retweet" etc.) to identify any possible entity. In short, the original tweet was sliced into various fragments by using POS tags, stop words, common words and punctuations as separators or delimiters. For example, if the content is “Samsung is holding a galaxy contest!”, two fragments will be generated for the content as follows: (samsung) | (galaxy contest).

2.3 Seed Words Generation

All the tweets extracted from samsungsg were subjected to data cleaning and preparation mentioned in the previous section. The process enables each tweet to be represented by the identified fragments or words and phrases. This set of data was further processed using term frequency analysis to obtain a list of seed words (which include “samsung”, “galaxy s iii”, “galaxy camera” etc.). The words in a phrase were joined by ‘_’ so that they could be identified as a single term but the ‘_’ was later filtered in all the matching processes.

These seed words were used to generate results for Fuzzy Keyword Match (see Section 2.4) and identify suitable topic numbers in the Twitter LDA method (see Section 2.5).

2.4 Fuzzy Keyword Match

It is not uncommon for Twitter users to use abbreviations or interjections or a different form of expression to represent similar terms. For example, “galaxy s iii” can be represented by “galaxy s 3”, which is understandable by a human but cannot be captured through a direct keyword match method. As such, a Fuzzy Keyword Match method using the seed words derived was implemented in this study.

The comparison here is based on a Dice coefficient string similarity score [17] using the following expression:

$$s = 2{*}n_t/(n_x+n_y) \tag{1}$$

where n_t is the number of characters found in both strings, n_x is the number of characters in string x and n_y is the number of characters in string y. For example, 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 is 2*4/(6+6) = 0.67. Each of the tweets of every follower is compared with the seed words and the highest score of any match is maintained as the *s* score of the follower.

2.5 Twitter LDA

Recently, LDA [12], a renowned generative probabilistic model for topic discovery, has been used in various social media studies [18][19]. 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. As such, we have adopted the implementation of Twitter LDA [18] for unsupervised topic discovery among all the followers.

As the volume of the tweets from all the followers in this study was within 200,000, a small number of topics (from 10-30, with an interval of 10) from Twitter LDA were used. We ran these topic models for 100 iterations of Gibbs sampling while keeping the other model parameters or Dirichlet priors constant: $\alpha = 0.5$, $\beta_{word} = 0.01$, $\beta_{background} = 0.01$ and $\gamma = 20$. Suitable topics were chosen automatically via comparison with the list of seed words. The result or the audience list identified by each topic model was a consolidation of 30 runs. The score assigned to each follower can be calculated using the following equation:

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

where n_m is the total number of matches and n_r is the total number of runs. If a particular follower is found in five runs then the t score assigned is 5/30 = 0.17.

2.6 The SVM

The SVM is a supervised learning approach for two- or multi-class classification, and has been used successfully in text categorisation [13]. 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 respectively. 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 [20].

In this work, the positive training data was generated using processed tweets from samsungsg, the selected account owner. The negative data was randomly generated from account owners of 10 different domains (online shopping deals, food, celebrities, parents, education, music, shopping, politics, Singapore news, traffic), which include ilovedealssg, hungrygowhere, joannepeh, kiasuparents, MOEsg, mtvasia, tiongbahruplaza, tocsg (TheOnlineCitizen), SGnews and sgdrivers respectively. These domains were chosen as they were the main topics discovered using Twitter LDA from the tweets of the followers of samsungsg. The respective account owners were selected as they were the popular Twitter accounts in Singapore according to online Twitter analytic tools such as wefollow.com.

The LibSVM implementation of RapidMiner [21] was used in this study and the sigmoid kernel type was selected as it produces higher precision prediction than other kernels, such as the radial basis function and polynomial.

A v score was assigned for each follower according to individual tweet classification based on the SVM. The v score can be generated using the following equation:

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

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. If a follower has tweeted two related tweets out of a total of 10 tweets, the v score assigned will be 0.2.

2.7 Construction of the HVSA Index

While it is possible to use the s, t and v scores individually as an index for segmenting and identifying a high-value social audience member, each method has its own strengths and limitations. It is therefore of interest to analyse if the combination of the various scores can generalise the identification task and help to improve the classification result.

An average value of scores from five methods, namely Fuzzy Keyword Match, Twitter LDA with 10, 20 and 30 topic models, and the SVM, was used in this study to generate the combined score. As there were three Twitter LDA methods considered, an analysis was done to assess if it would be feasible to use just one of the Twitter LDA methods in developing a representative HVSA index.

The threshold used for the HVSA index is based on the ranking of the scores generated. For example, the HVSA index at top 100 represents the average score value of the top 100 scorers according to the methods of scoring. This top scorer segmentation approach has been adopted as it resembles a real world scenario where a company will more likely be interested in identifying the top n potential customers in an attempt to maximise the use of their marketing resources.

3 Experiments and Results

The results obtained from the various methods were compared with a random annotated sample of the followers of samsungsg. The contents of a total of 300 followers (which were randomly sampled) were annotated manually as either a potential high-value social audience according to the content shared by the account owner or not a target audience. This set of data was used in the evaluation of the various methods and detailed analyses can be found in Sections 3.1, 3.2 and 3.3.

3.1 Results of Various Methods

To compare the various methods, Receiver Operating Characteristic (ROC) curves, as shown in Fig. 2, are plotted for all the results using the various scores derived (Fuzzy Keyword Match uses the s score, Twitter LDA methods use the t score, and the SVM uses the v score). It is observed that Fuzzy Keyword Match has obtained the largest area under the curve, followed by the Twitter LDA topic modelling methods and the SVM.

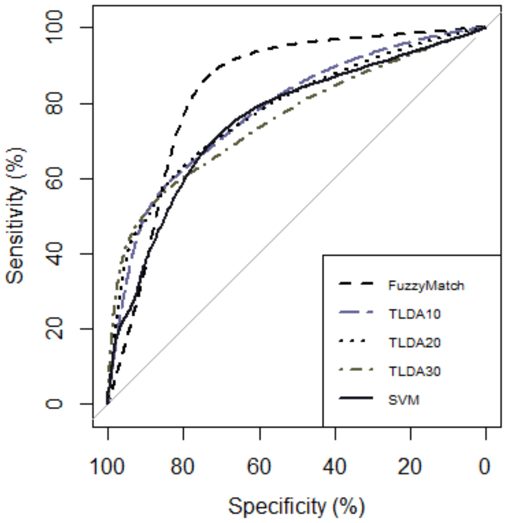

Fig. 2. ROC curves of various methods

The corresponding Area Under Curve (AUC) value can be found in Table 1. It is interesting to note that the Twitter LDA methods have generated similar AUC values irrespective of the number of topics used. It is therefore worth analysing whether it is possible to use just one of the Twitter LDA methods to represent the HVSA index instead of all three methods.

Table 1. AUC for various methods

Methods	AUC
Fuzzy Keyword Match	0.88
Twitter LDA 10 topics	0.81
Twitter LDA 20 topics	0.81
Twitter LDA 30 topics	0.78
SVM	0.80

3.1.1 Results of Twitter LDA

As shown in Fig. 2 and Table 1, the various Twitter LDA methods have achieved similar trends and results. A further analysis was done and the group of audience members identified with topic numbers greater than 30 remained the same. Hence, only the results of topic models from 10, 20 and 30 are included in this paper.

Table 2 presents some sample topic groups and their topical words. The table shows that using seed words derived from the account owner can identify relevant contents from the list of followers.

Table 2. Sample topic groups and their topical words (ID is the topic group id)

Models	IDs	Top topical words
Twitter LDA 10 topics	3	google, android, apps, mobile, galaxy, tablet
	4	samsung, galaxy, mobile, phone, android, tv, camera, smartphone
Twitter LDA 20 topics	8	galaxy, samsung, android, phone, mobile, apps, smartphone
	17	samsung, galaxy, app, tablet
	19	samsung, tv, led, mobile, smart, phone, laptop
Twitter LDA 30 topics	3	samsung, galaxy
	18	samsung, galaxy, android, google, app, phone, mobile, tablet, smartphone
	25	samsung, tv, led, camera, lcd, smart, hd

3.1.2 Results of the SVM

The 10-fold cross validation of the training data yields an accuracy of 88%, with class precision and recall as presented in Table 3. However, when we applied the model on the testing data of the followers' tweets, the denominator (or the normalisation process) used in Eq. (3) plays an important role as further investigation has shown that using an average value of all the tweets only yields an AUC value of 0.66 [22] instead of 0.80 (as shown in Table 1). In this study, the total number of tweets was used to normalise the score instead of an average value of all the tweets due to the fact that the resulted score is more capable in representing the true interest of a follower. For example, if follower1 has tweeted two related tweets out of a total of 10 tweets, the *e* score assigned will be 0.2. While the *e* score for follower2 is 0.02 if only two related tweets are classified as positive out of a total of 100 tweets. This is in contrast to using an average value, as both follower1 and follower2 will be assigned the same *e* score, which may not fully represent the interests of the followers.

Table 3. SVM 10-fold cross validation results

	True samsungsg	True others	Class precision
Predicted samsungsg	165	13	92.7%
Predicted others	35	187	84.2%
Class recall	82.5%	93.5%	

3.2 Analysis of HVSA Index Construction

As mentioned in Section 2.7, the construction of the HVSA index is based on the ranking of average scores generated from various methods. In other words, the segmentation is done such that the top 100 threshold represents the top 100 scorers. The percentage of match with the annotated data together with the *s*, *t* and *v* scores under different segmentations can be found in Table 4. As expected, the *s* score has a better coverage and hits 92% at the top 1000 threshold. However, it is worthwhile to note that it has not performed as well as the Twitter LDA methods at the top 250 mark.

In general, all three Twitter LDA methods have the similar trend but Twitter LDA with 20 topic models is able to cover all 100% of the annotated dataset at the top 2000 threshold. The SVM (i.e., the *v* score) does not cover the percentage match as well as the others but as it is a machine learning approach, it has the potential to be better with more training data. Furthermore, it is possible that the SVM is more selective as it is not directly using any keyword matching approach. Besides that, the *t* scores from the Twitter LDA methods are generally more evenly distributed while the *s* score of Fuzzy Keyword Match has a higher value and the *v* score of the SVM has a lower value as shown in Table 4.

Table 4. Comparison of the various scores and their percentage of match with the annotated data (where *t10* score represents the *t* score from Twitter LDA 10 topics and *t20* represents the *t* score from Twitter LDA 20 topics and so on)

Top	*s* score	% match	*t10* score	% match	*t20* score	% match	*t30* score	% match	*v* score	% match
100	1.0	27%	1.0	33%	1.0	29%	0.97	24%	0.38	8%
250	1.0	27%	0.97	40%	0.97	32%	0.87	35%	0.25	22%
500	0.83	84%	0.8	54%	0.6	51%	0.53	52%	0.17	40%
750	0.83	84%	0.5	65%	0.33	67%	0.3	63%	0.13	60%
1000	0.75	92%	0.37	75%	0.2	79%	0.2	70%	0.11	70%
1500	0.6	98%	0.2	92%	0.1	87%	0.1	86%	0.07	87%
2000	0.57	100%	0.03	97%	0.03	100%	0.03	95%	0.01	90%

Table 5. Comparison of various HVSA index constructions

Top	Five-method average score	% match	Three-method average score	% match
100	0.82	16%	0.72	16%
250	0.75	35%	0.65	27%
500	0.56	54%	0.51	59%
750	0.39	67%	0.41	78%
1000	0.31	86%	0.36	86%
1500	0.23	95%	0.29	97%
2000	0.18	100%	0.23	100%

As the Twitter LDA 20 topics model has performed well, as shown in Table 4, it is used as a representative of the Twitter LDA methods for the HVSA index construction. In order to evaluate if it is sufficient to use the score from just one of the topic models (instead of all three), the average score from all the five methods considered in Table 4 and the average score from three of the methods, namely Fuzzy Keyword Match, Twitter LDA with 20 topic models and the SVM, are presented in Table 5. In general, the three-method average has a better coverage as compared to the five-method one even though it has a lower percentage match at the top 250 threshold.

3.3 Percentage of High-Value Social Audience Identified through Analysis of Top Scorers

By using the three-method average score (as shown in Table 5) as the HVSA index, each of the followers has a corresponding HVSA index calculated according to the average of the *s*, *t20* and *v* scores. These indices are compared with the top scorer of the respective segmentations and followers with indices falling within the range will be considered as a match. The percentages of match based on the respective HVSA index for all followers of samsungsg as well as the annotated followers are listed in Table 6. Here, the HVSA index generated for top 100 is the value assigned to the top 100 scorers based on the index construction approach mentioned in Section 3.2. With a HVSA index of 0.36 or at the threshold of top 1000 scorers, 86% of the annotated followers are identified, which covers 28% of all the followers. All the annotated followers or the potential target audience actually have HVSA indices greater than 0.23, while this value covers half of all the followers. In short, by introducing the HVSA index, ranking of the followers and selection of the target audience can be done and potentially be more effective as compared to randomly selecting anyone from all the followers.

Table 6. Top scorers with the corresponding percentage for all the followers and the annotated followers

Top	HVSA Index	% within all the followers	% in the annotated followers
100	0.72	3%	16%
250	0.65	7%	27%
500	0.51	14%	59%
750	0.41	21%	78%
1000	0.36	28%	86%
1500	0.29	42%	97%
2000	0.23	56%	100%

4 Discussion

It is interesting to observe from the results that, the proposed HVSA index is able to identify the high-value social audience from the annotated random users. While the percentage of matches in the annotated data can be used as a guide to assess the potential usefulness of this HVSA index, it is also worthwhile to analyse the detailed contents of some of the followers whose scores from various methods are not in agreement. Some examples are listed in Table 7.

From the table, we can see that `follower1` has consistent high scores for four of the five methods (except the SVM). A closer look at the tweets shows that the follower shares mostly technology and mobile news with tweets like `"RT @ZDNet: Samsung announces Galaxy S Wi-Fi 3.6"` and `"Google Chrome has 70 million Active Users"`. Having an HVSA index of 0.49 put the follower in the

Top 500 to Top 750 range making them likely to be considered a member of a high-value audience. In contrast, the *v* score of 0.06 derived from the SVM method alone does not reveal this.

Another scenario of inconsistency is when any two scores, for example, *s* and *v* scores (or the scores from Fuzzy Keyword Match and the SVM respectively), are high but the three Twitter LDA methods are not. `follower2` falls under this scenario but the HVSA index indicates that this follower is highly likely to be a target audience member. A detailed study on the contents shared by `follower2` indeed shows a tweet asking samsungsg about the Samsung galaxy S3 workshop.

Although Fuzzy Keyword Match has performed well, the method seems to have consistently higher scores than the other methods. This may lead to the identification of false positives due to the higher value assigned. In Table 7, `follower3` was scored badly by all the other methods except for Fuzzy Keyword Match (the *s* score has the value of 1.0). A detailed investigation on the user's tweets reveals that 98 tweets extracted were mostly about school studies and daily activities, even though there were also two tweets mentioning about the phone: `"My phone is useless now after updating my phone!!!"` and `"resetting my phone :( "`. This follower was in fact a non-target audience member (as per the manual annotation). As such, it is worth combining the various scores in deriving a more suitable score or index for identifying the high-value social audience. The HVSA index for `follower3` is 0.3, which falls within the range of top 1000 to top 1500. In actual fact, as the threshold of top 1500 is 0.29, this follower is likely to be in the upper range and less likely to be identified as a target audience member as compared to having a single *s* score of perfect 1.0.

As discussed above, there is some benefit in using the HVSA index over individual scores, as each method has strengths and weaknesses. By combining the scores, a more general index could be derived, that would be more practical and useful in real-world applications.

While the above discussion is about the annotated target audience, we have also done some detailed study on the top scorers of the annotated non-target audience. A total of six followers having a HVSA index between 0.5 and 0.6 were identified. A close look at each of the followers shows that three of them are indeed target audience members who had shared similar contents as the account owner, samsungsg. These three followers are mainly technology and mobile news Twitter users while the other three are not directly related. One of the latter had shared mainly contents related to iPhone/iPad, while the other two mentioned Samsung in some of their tweets but the tweets were really about doing business and launching a complaint with Samsung.

Although a human annotator is preferred most of the time, it is a challenge to annotate tens of hundreds of tweets, and mistakes are inevitable. On the other hand, the HVSA index can be handy as a first cut to identify the high-value social audience from a huge list of followers without the need to manually annotate each of them. In order to increase accuracy, verification through a human can be applied on followers with inconsistent scores, which definitely aids in minimising the annotation effort. Engagement done through this approach is definitely better than selecting followers randomly or manually selecting them based on keywords.

Table 7. Interesting followers identified. The higher scores of each user are bolded. The *s* score is generated by Fuzzy Keyword Match, *t10* score is generated by Twitter LDA 10 topics, *t20* score is generated by Twitter LDA 20 topics, *t30* score is generated by Twitter LDA 30 topics and *v* score is generated by the SVM.

Twitter name	*s* score	*t10* score	*t20* score	*t30* score	*v* score	HVSA index
follower1	**0.83**	**0.97**	**1.0**	**0.8**	0.06	0.49
follower2	**1.0**	0.53	0.43	0.53	**0.33**	0.52
follower3	**1.0**	0.37	0.17	0.0	0.03	0.3

5 Conclusion and Future Work

In this study, we have constructed a High-Value Social Audience (or HVSA in short) index from various text mining methods to identify the high-value social audience from a list of followers using the contents of a Twitter account owner, samsungsg. It is assumed that those who have tweeted similar contents are more likely to be interested in the owner's tweets, compared to those who have not been sharing similar contents.

Our results show that the HVSA index is a better indicator than individual scores from the various methods, as the index is an aggregate of those scores and hence it is capable of combining all the findings and providing a more generalised outcome. It is more practical and possibly more useful for a real-world marketing application.

While currently the index is the average value of several methods, other index construction approaches considering the precision or recall values may be incorporated to derive a more robust indicator. It should be noted that the index is developed as a guide for customer segmentation in the application area of targeted marketing. This means any improvement over mass marketing is going to be beneficial for business companies.

We have used samsungsg as a case study in this paper. It has been shown that contents extracted from the account owner can be used to identify the target audience. For future work, we plan to extend it to include other account owners to verify if the observation is consistent across Twitter or if there is any pattern observed for different types of Twitter accounts. For example, a more generic account on parent groups or current affairs may have contents that are more diverse and conceptual and may not work well with keyword-based matching methods like Fuzzy Keyword Match. As such, a more sophisticated feature generation method based on domain-specific and common-sense knowledge may be required to enrich the bag of words with new, more informative features.

References

1. How Many People Use Facebook, Twitter and 415 of the Top Social Media, Apps & Tools (updated (March 2014), `http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/#.Uz0f4Vc4t5E`
2. Unlocking the power of social media | IAB UK, `http://www.iabuk.net/blog/unlocking-the-power-of-social-media`
3. 2013 Fortune 500 - UMass Dartmouth, `http://www.umassd.edu/cmr/socialmediaresearch/2013fortune500/`

4. Mo, J., Kiang, M.Y., Zou, P., Li, Y.: A two-stage clustering approach for multi-region segmentation. Expert Systems with Applications 37, 7120–7131 (2010)
5. Namvar, M., Khakabimamaghani, S., Gholamian, M.R.: An approach to optimised customer segmentation and profiling using RFM, LTV, and demographic features. International Journal of Electronic Customer Relationship Management 5, 220–235 (2011)
6. Greenberg, P.: CRM at the Speed of Light: Social CRM 2.0 Strategies, Tools, and Techniques for Engaging Your Customers. McGraw-Hill Osborne Media (2009)
7. Malthouse, E.C., Haenlein, M., Skiera, B., Wege, E., Zhang, M.: Managing customer relationships in the social media era: introducing the social CRM house. Journal of Interactive Marketing 27, 270–280 (2013)
8. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 251–260. ACM (2010)
9. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 110, 5802–5805 (2013)
10. How Ebay Uses Twitter, Smartphones and Tablets to Snap Up Shoppers, http://www.ibtimes.co.uk/how-ebay-uses-twitter-smartphones-tablets-snap-shoppers-1443441
11. Zhang, Y., Pennacchiotti, M.: Predicting purchase behaviors from social media. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1521–1532 (2013)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
13. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer (1998)
14. Using the Twitter Search API | Twitter Developers, https://dev.twitter.com/docs/using-search
15. Nakatani, S.: Language-detection - Language Detection Library for Java - Google Project Hosting, http://code.google.com/p/language-detection/
16. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, vol. 13 (2000)
17. Kondrak, G., Marcu, D., Knight, K.: Cognates can improve statistical translation models. In: 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, vol. 2 (2003)
18. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. Advances in Information Retrieval, pp. 338–349. Springer (2011)
19. Yang, M.-C., Rim, H.-C.: Identifying interesting Twitter contents using topical analysis. Expert Systems with Applications 41, 4330–4336 (2014)
20. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
21. Predictive Analytics, Data Mining, Self-service, Open source - RapidMiner, http://rapidminer.com/
22. Lo, S.L., Cornforth, D., Chiong, R.: Identifying the high-value social audience from Twitter through text-mining methods. In: Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolution Systems, vol. 1, pp. 325–339 (2014)