10-2014

Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on twitter

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

David CORNFORTH

Raymond CHIONG

Effects of Training Datasets on Both the Extreme Learning Machine and Support Vector Machine for Target Audience Identification on Twitter

Siaw Ling Lo, David Cornforth, and Raymond Chiong

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

Abstract. The ability to identify or predict a target audience from the increasingly crowded social space will provide a company some competitive advantage over other companies. In this paper, we analyze various training datasets, which include Twitter contents of an account owner and its list of followers, using features generated in different ways for two machine learning approaches - the Extreme Learning Machine (ELM) and Support Vector Machine (SVM). Various configurations of the ELM and SVM have been evaluated. The results indicate that training datasets using features generated from the owner tweets achieve the best performance, relative to other feature sets. This finding is important and may aid researchers in developing a classifier that is capable of identifying a specific group of target audience members. This will assist the account owner to spend resources more effectively, by sending offers to the right audience, and hence maximize marketing efficiency and improve the return on investment.

Keywords: Extreme learning machine, Support vector machine, Machine learning, Target audience, Twitter, Social media.

1 Introduction

With the prevalence of social media and openness of information sharing, the ability to analyze the contents of public social media posts and to uncover the underlying insights is valuable to any organization. Doing business on social media is becoming common, when one considers that 77% of the Fortune 500 companies have active Twitter accounts and 70% have active Facebook accounts [1]. It is understandable that most companies are making efforts to engage their customers in more than one social platform, as it can be rewarding to reach out to potential customers from the huge user base of over 1.28 billion from both Twitter and Facebook [2].

In view of the increasingly crowded social space, it is no longer feasible for a company to depend on gimmicks (such as incentive referrals) that may only provide short-term gain. With the growing "sophistication" of social media users, approaches like mass marketing may not justify the effort and amount of money spent. Furthermore, there is a thin line between broadcasting a general message and spamming, so instead

J. Cao et al. (eds.), *Proceedings of ELM-2014 Volume 1*,
Proceedings in Adaptation, Learning and Optimization 3, DOI: 10.1007/978-3-319-14063-6_35

of attracting a greater audience, there is a high risk of alienating and therefore losing current customers. Hence, it makes sense to identify a target audience in order to maximize marketing efficiency and improve the return of investment (ROI).

While there are many guidelines or tips on the web that suggest how to find a target audience on social media, most of these concentrate on searching specific keywords related to products or brands. This list of keywords is usually prepared by domain experts, and there is a need to ensure that the keywords are "up-to-date" or stay relevant, due to the dynamism of the business world. Furthermore, deciding which keywords to use may not be obvious to a non-expert and this may lead to inaccurate information extraction and hence a misunderstood market analysis. On top of this, there is a need to manually consolidate the list of members of a social audience found and to ensure that the contents shared by the audience match the keywords.

Prior work [3][4] has proposed various approaches such as translating both social networks and semantic information into Resource Description Framework (RDF) formats and using RDF methods for correlation, or making use of semantic tagging to correlate the current social tagging approach to make sense of the social media data. These approaches, however, require additional efforts of translating and tagging of current social media data, which can be a daunting task considering the huge amount of data and the possible manual procedure involved.

While the ability to predict a target audience from a list of Twitter followers will be beneficial to any company or organization, limited studies have investigated the effects of different training datasets used in supervised machine learning approaches for this purpose. The most relevant work by Yang et el. [5] used temporal effects of Twitter contents and a list of category-specific keywords to classify users' interests in the sports and politics domains. Both Term Frequency-Inverse Document Frequency (TF-IDF) [6] and Latent Dirichlet Allocation (LDA) [7] have been used by Yang and co-workers to generate features in two classification approaches – Naïve Bayes (NB) [8] and the Support Vector Machine (SVM) [9]. In comparison, our work focuses on predicting the target audience from a list of followers of a Twitter account owner, instead of analyzing a domain specific Twitter user interest. In addition, we extract a set of seed words using the contents of the owner's tweets, which reflects the key topics or terms at a specific point of time. This can aid in generating various training datasets, while eliminating the need for manual tagging.

It is well known that one of the biggest challenges of using a supervised machine learning approach is the constructing of its training dataset. Due to the vast amount and diverse nature of the followers' tweets, it is not feasible to manually annotate the tweets for training the machine learning model. As such, we reasoned that tweets from an account owner can be used to build a positive training dataset as the group of followers who are tweeting similar contents (within a similar period of time) are more likely to comprise the target audience compared to others who are not sharing similar contents. This saves us from the need to manually annotate the vast amount of tweets from the followers and is more practical if the approach is to be adopted in a real-world application.

Machine learning approaches, especially the SVM, have been used in various text categorization tasks [9] and are found to have obtained better performances compared

to other methods [10][11]. As we are analyzing the textual content of tweets, it is of interest to study how the SVM would perform in predicting the target audience from a list of followers. Recently, biologically inspired Natural Language Processing (NLP) [12] has started to gain popularity and a new approach known as the Extreme Learning Machine (ELM) [13] has achieved good results in unstructured text analysis. This research will therefore use both the SVM and ELM for target audience classification. As the performance of a machine learning approach depends heavily on its training dataset, it is important to construct features that can represent the dataset in an appropriate manner. It is logical to consider the suitability of the features according to specific domain knowledge and human ingenuity. Hence, we derive our training datasets from the contents of an owner's tweets. The hypothesis here is based on the idea that the owner's tweets will contain the necessary information or features that the followers are interested and so they choose and take action to follow the account owner.

In this paper, our main objective is to investigate the effects of using different training datasets on both the ELM and SVM in order to make use of available resources to predict and identify the target audience without utilizing a considerable amount of human annotation effort. There are three types of training datasets: 1) tweets from the owner, 2) tweets from followers clustered using statistical topic modeling - Twitter LDA [14], and 3) tweets from followers generated through fuzzy matching using a list of seed words extracted from the owner's tweets.

The major contributions of this work are as follows:

- To the best of our knowledge, our work is the first attempt to predict and identify a target audience from a list of followers (of an account owner) on Twitter using the ELM with minimum manual annotation required.
- From our observation of the results, features generated using the content of the owner's tweets in the training dataset are more useful than training datasets that use the followers' tweets.
- We find that it is essential to remove all the duplicates after the data cleaning process, in order to improve the classification result.

2 Related Work

The aims of any business are to increase profit, build a long lasting brand name, and to grow the customer base or engage current customers. It is therefore essential for a company to understand the needs and behavior of its customers. This understanding can be achieved through different means and at different levels of detail. Most companies segment their customers according to their traits and behavior so that marketing activities are targeted and measured according to the segmentation.

However, this kind of 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 [15, 16]. While CRM or organizational transaction data can be coupled with

geographical data to obtain additional information, the segmentation remains limited to a company's internal system and does not leverage on the sharing and activities on social media where customers tend to reveal more about themselves such as personal preferences and perception of brands.

There have been efforts in deriving or estimating demographics information [17, 18] from available social media data, but this set of information may not be suitable to be used directly in targeted marketing, as temporal effects and types of products are usually not considered. Besides that, demographic attributes such as age, gender and residence areas may not be updated and hence may result in a misled conclusion. Recently, eBay has expressed that, due to the viral campaigns and major social media activities, marketing and advertising strategies are evolving. Targeting specific demographics through segmentation, although this still has value, is being superseded by content-based approaches: eBay is focusing on "connecting people with the things they need and love, whoever they are" [19]. In other words, contents shared by individuals are more important than demographics for predicting the target audience on social media. Other research has also shown that using Facebook categories, such as likes and n-grams, for predicting purchase behavior from social media is better than using demographic features shared on Facebook [20]. 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.

There are many approaches in understanding the preferences of Twitter users, which can provide important opportunities for businesses to improve their services (such as through targeted marketing or personalized services). The majority of these approaches focus on classifying Twitter users using the textual features (e.g., contents of the tweets) [21] or network features (e.g., follower/followee network) [22]. However, there is also work in which the researchers have adopted various sociolinguistic features such as emoticons and character repetition [22], and they used a SVM to classify latent attributes such as gender, age, regional origin and political orientation. Ikeda et al. [23] developed 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 are detected by clustering their followers and followees.

As most of the Twitter users' basic demographic information (e.g., gender, age) is unknown or incomplete (as compared to Facebook), Yang et al. [5] examined the temporal effect of Twitter contents or tweets in classifying users' interests. Instead of using tweets directly, temporal information is derived from the word usage within the streams to boost the accuracy of the classification. Both binary- and multi-class classifications have been tested and found to outperform other methods within the sports and politics domains. Another approach by Hong et al. [24] modeled a user's interest and behavior by focusing on retweet actions in Twitter, which can be used to model user decisions and user-generated contents simultaneously. 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. Hence, it is essential to extract relevant information from tweets for user profiling tasks. Michelson and Macskassy [25] presented work in discovering topics of interest by examining the entities in

tweets. A "topic profile" is then developed to characterize the users. Besides that, statistical techniques that extract term- and concept- based user profiles are used to analyze customers' conversational data to provide insights on a user's interest so that commercial services can use these profiles for targeted marketing [26].

3 Methods

The focus of this work is to understand the effect of using different training datasets on predicting the target audience from a list of followers via two machine learning methods, namely the ELM and SVM.

3.1 Data Collection

We use the Twitter Search API [27] for our data collection. As the API is constantly evolving with different rate limiting settings, our data gathering is done through a scheduled program that requests a set of data for a given query. The subject or brand selected for this research is *Samsung Singapore* or "samsungsg" (its Twitter username). At the time of data collection, there were 3,727 samsungsg followers. In order to analyze the content of the account owner's tweets, the last 200 tweets by samsungsg have been extracted. The time of tweets ranges from 2 Nov 2012 to 3 Apr 2013. For each of the followers, the API is used to extract their tweets, giving a total of 187,746 records, and 2,449 unique users having at least 5 tweets in their past 100 tweets of the same period. We reasoned that those with fewer than 5 tweets were inactive in Twitter, as it implied that these user were tweeting an average of less than one tweet in a month (since the period was of 6 months).

3.2 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 are removed using the Language Detection Library for Java [28];
- URLs, any Twitter's username found in the content (which is in the format of @username) and hashtags (with the # symbol) are removed;
- Each tweet is pre-processed to lower case.

As tweets are usually informal and short (up to 140 characters), abbreviation and misspelling 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. Due to this, we derive an approach called Entities Identification, which uses Part-of-Speech (POS) [29] tags to differentiate the type of words. All the single nouns are identified as possible entities. If the tag of the first fragment detected is 'N' or 'J' and the consecutive word(s) is of the 'N' type, these words will be extracted as phrases. This approach is 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 is 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).

3.3 Extreme Learning Machine (ELM)

The ELM [30] is a single-hidden layer feedforward neural networks (SLFN) where all the node parameters in the hidden layer are randomly generated without tuning. Through the replacement of a computationally costly procedure of training the hidden layer by using random initialization, the method is proven to have both universal approximation and classification capabilities [31][32].

Consider a set of N distinct samples (x_i, y_i) with $x_i \in \Re^D$ and $y_i \in \Re^d$. An ELM with K hidden neurons is modeled as

$$\sum_{k=1}^{K} \beta_k \phi(w_k x_i + b_k), \quad i \in [1, N] \tag{1}$$

where ϕ is the matrix activation function, w the input weights, b the biases and β the output weights.

A MATLAB implementation of ELM from *http://www.ntu.edu.sg/home/egbhuang/elm_random_hidden_nodes.html* is adopted in this study. A sigmoidal function is used as the activation function instead of the alternatives, as it has performed better on the various training datasets mentioned in Section 3.6. A range of numbers from 50-400 are used as the hidden neuron parameters, with an interval of 50.

As the number of tweets shared by each follower is different, an e score is calculated by aggregating the classification results from each individual tweet of each follower's tweet set. The final assignment of the e score is based on the following representation:

$$e = n_p / n_t \tag{2}$$

where n_p is the total number of tweets that are classified as positive by the ELM and n_t is the total number of tweets shared by the follower. If 5 tweets out of a total of 50 tweets of a particular follower are classified as positive, then the e score assigned is 5/50 = 0.01. The total number of tweets is used to normalize the score instead of an average value of all the tweets. This is due to the fact that the resulted score is more capable of representing the true interest of the follower. For example, if follower1 tweeted 2 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 2 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 that may not fully represent the interests of the followers.

3.4 Support Vector Machine (SVM)

The SVM is a supervised learning approach for two- or multi-class classification and it has been used successfully in many applications, including text categorization [9]. It separates a given known set of {+1, -1} labeled 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 [33].

Consider a set of N distinct samples (x_i, y_i) with $x_i \in \Re^D$ and $y_i \in \Re^d$. An SVM is modeled as

$$\sum_i \alpha_i K(x, x_i) + b, \quad i \in [1, N] \tag{3}$$

where $K(x, x_i)$ is the kernel function, and α and b are the parameter and threshold of the SVM respectively.

The LibSVM implementation of RapidMiner [34] is used in this study, and the sigmoid kernel type is selected as it produces higher precision prediction than other kernels, such as RBF (Radial Basis Function) and polynomial.

Similar to the e score specified in Section 3.3, a v score is assigned for each follower according to individual tweet classification based on the SVM. The v score is generated using the following formula:

$$v = n_s/n_t \tag{4}$$

where n_s is the number of tweets that are classified as positive by the SVM and n_a is the total number of tweets shared by the follower.

3.5 Performance Metrics

The typical accuracy metric in statistical analysis of binary classification, which takes into consideration the true positive (TP) and true negative (TN), does not reflect the performance of a classifier well [35]. Therefore, we have used the precision, recall and F1 score as performance metrics when comparing the ELM and SVM.

The formulas of precision, recall and F1 score are as follows:

$$precision = TP/(TP + FP) \tag{5}$$

$$recall = TP/(TP + FN) \tag{6}$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

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

3.6 Generation of Training Datasets

As our main intention is to find an approach to predict the target audience from a list of followers without the need to manually annotate the vast amount of tweet contents for training purposes, we have designed the following three procedures:

i. Using tweets from the account owner (which logically should be tweeting contents that will attract followers of similar interest) as the positive dataset and tweets of account owners from other domains as the negative dataset;
ii. Using an unsupervised topic modeling approach to cluster relevant tweets from all followers as the positive dataset and other clusters as the negative dataset;
iii. Using the Fuzzy match approach with seed words extracted from tweets of the account owner to identify relevant tweets from all the followers. Those tweets that are matched with a certain threshold are assigned as the positive dataset and those below the threshold are assigned as the negative dataset.

Details of the various types of training datasets and their corresponding number of features can be found in Table 1.

Table 1. Types of training datasets and the number of features

Training datasets	Size of the datasets	Number of features	Notation
Tweets of owners	200 positive and 200 negative	245	owner
Tweets of followers generated using Twitter LDA	13,989 positive and 99,831 negative (7 sets of training datasets are created)	397	follower TLDA
Tweets of followers generated using Fuzzy match with seed words extracted from the owner	13,989 positive and 99,831 negative (7 sets of training datasets are created)	38	follower FV

3.6.1 Features Generated from the Owner's Tweets

The positive dataset is generated using processed tweets from the account owner (i.e., samsungsg). The negative dataset is randomly generated from account owners of 10 different domains, which include ilovedealssg (online shopping deals), hungrygowhere (food), joannepeh (celebrity), kiasuparents (parents), MOEsg (education), mtvasia (music), tiongbahruplaza (shopping), tocsg (TheOnlineCitizen/politics), SGnews (Singapore news) and sgdrivers (news on traffic) respectively. These domains are chosen as they represent the main topics discovered based on the analysis of Twitter LDA using the list of tweets from all the followers. The respective account owners are selected as they are the popular Twitter accounts in Singapore according to online Twitter analytic tools such as wefollow.com.

As all the tweets have been cleaned and preprocessed (see Section 3.2), only word stemming using Porter [36] is done on the tweets before forming a term frequency word vector. A total of 245 features are identified and used in creating both the training and testing feature vectors.

3.6.2 Seed Words Generation

In order to minimize the need to annotate the huge amount of followers' tweets for classification, seed words are extracted from the owner's tweets to assist in identifying relevant topic clusters in the unsupervised topic modeling approach (i.e., Twitter LDA, see Section 3.6.3) as well as the Fuzzy match approach (see Section 3.6.4).

All the tweets extracted from samsungsg are subjected to the data cleaning and preparation process described in Section 3.2. Each tweet is now represented by the identified fragments or words and phrases. This set of data is 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 are joined by '_' so that they can be identified as a single term but the '_' is filtered in all the matching processes. A total of 38 words and phrases are identified.

3.6.3 Features Generated Using Twitter Latent Dirichlet Allocation (LDA)

LDA, a renowned generative probabilistic model for topic discovery, has recently been used in various social media studies [14][37]. 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 [14] for unsupervised topic discovery among all the followers.

As the volume of the tweet set from all the followers is within 200,000 tweets, only a small number of topics from Twitter LDA have been used. Specifically, we have used five topic models from 10 to 50 (with an interval of 10) in this study. We ran these five different 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 are chosen automatically via comparison with the list of seed words.

The list of topic words under the selected topics are checked for duplication and a total of 397 words are identified for creation of training and testing datasets.

3.6.4 Features Generated Using Fuzzy Match with Seed Words Extracted from the Owner's Tweets

3.6.4.1 Fuzzy Match

It is not uncommon for Twitter users to use abbreviations or interjections or different forms 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 by direct keyword match. As such, fuzzy matching based on the seed words derived is implemented.

The comparison here is based on a Dice coefficient string similarity score [38] using the following expression

$$s = 2*n_c/(n_x+n_y) \quad (8)$$

where n_c 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, to calculate the 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) = 0.667.

3.6.4.2 Features Generated from Fuzzy Match

As the Fuzzy match method is dependent on the list of seed words extracted from the owner's tweets, the total number of features for it is the same as the number of seed words, which is 38. Each of the tweets from each follower is compared with every seed word using Fuzzy match. The highest similarity score of each seed word match for the tweet is used to create the value of the feature for that seed word of the tweet.

3.7 Generation of Testing Datasets

In order to assess the performance of both the ELM and SVM, the contents of a total of 300 followers (which were randomly sampled) were annotated manually as either a potential target audience or not a target audience based on the contents shared by the account owner.

Even though the original tweet contents from the annotated followers were mostly different, the contents of the testing dataset after the data cleaning and preparation process (see Section 3.2) resulted in a fair amount of duplication (as shown in Table 2). Hence, it will be of interest to analyze if the duplication in the testing dataset has any effect on the performance of the classifiers.

Table 2. Types of testing datasets

Testing datasets	Size of the datasets	Notation
All tweets from annotated followers	21,297	nil
Unique tweets from annotated followers	13,550	noDup

4 Experiments and Results

4.1 Training Accuracy and *e* Scores of Various ELM Configurations

As all the hidden node parameters are randomly generated in the ELM, 10 runs using different ranges of hidden nodes or neuron numbers have been carried out. It is observed

that neuron numbers within the range of 150 and 250 produced better results. The average training accuracy and time with experiments on 150, 200 and 250 neuron numbers are listed in Table 3.

Table 3. Training accuracy of various ELM configurations

Training datasets	Neuron numbers	Training accuracy*	Training time (s)*
ELM_owner	150	0.96	0.07
	200	0.99	0.13
	250	0.99	0.19
ELM_followerTLDA	150	0.67	2.38
	200	0.68	4.05
	250	0.69	5.42
ELM_followerFV	150	0.59	2.02
	200	0.60	3.49
	250	0.60	5.36

*The results are based on the average of 10 runs

The training model using 250 neuron numbers has consistently performed well compared to other configurations and hence we used it for testing the two different types of testing datasets generated using the tweets of the 300 randomly annotated followers. As indicated in Table 4, there are two testing datasets for each training model - the complete set and the no-duplicate set. An *e* score is generated for each follower and the top 10 and top 30 *e* scores are listed in the table. These two sets of scores have been selected to assess how well the ELM performs in identifying and predicting a target audience. It is beneficial to know how many of the top 10 followers as predicted by the ELM are true target audience members in addition to looking at scores from performance metrics such as precision, recall and F1. Detailed results of these can be found in Section 4.3 and 4.4.

Table 4. *e* scores of various ELM configurations

Training – Testing	Top 10 *e* scores	Top 30 *e* scores
ELM_owner	0.50	0.33
ELM_owner_noDup	0.35	0.21
ELM_followerTLDA	0.68	0.43
ELM_followerTLDA_noDup	0.60	0.41
ELM_followerFV	0.72	0.56
ELM_followerFV_noDup	0.70	0.58

4.2 Training Accuracy and *v* Scores of Various SVM Configurations

The 10 fold cross-validation result for the SVM has yielded an accuracy of 0.88 when owner contents are used as the training dataset. As shown in Table 5, the other training datasets are not doing as well as the SVM_owner training dataset.

Table 5. Training accuracy of various SVM configurations

Training datasets	Training accuracy*
SVM_owner	0.88
SVM_followerTLDA	0.70
SVM_followerFV	0.57

*The results are from 10 fold cross-validation

Similar to the ELM, the top 10 and top 30 *v* scores were generated for the assessment of various SVM configurations. It is interesting to observe that there is an increasing trend for the *v* score from using the owner tweets as the training dataset to followerTLDA and finally followerFV (see Table 6). This observation is consistent with the results obtained for various ELM configurations (as shown in Table 4) and it implies that the identification of target audience becomes less specific in other training datasets as compared to the owner training dataset. Besides that, the scores (both the *e* and *v* scores) are lower when the no-duplicate testing dataset is used.

Table 6. *v* scores of various SVM configurations

Training – Testing	Top 10 *v* scores	Top 30 *v* scores
SVM_owner	0.42	0.24
SVM_owner_noDup	0.33	0.18
SVM_followerTLDA	0.69	0.42
SVM_followerTLDA_noDup	0.71	0.39
SVM_followerFV	0.80	0.63
SVM_followerFV_noDup	0.75	0.60

4.3 Results of Using Top 10 Scores as Cut Off

Table 7 and Table 8 show the results of using top 10 scores as cut off for the ELM and SVM respectively. In general, the numbers of true positive (TP) identified decrease from using the owner training dataset to the followerFV dataset. However, it is worth highlighting that using the no-duplicate testing dataset has yielded better results compared to the complete set of test data here.

Table 7. Results of the ELM (based on the average of 10 runs, using 250 neurons) – top 10 score cut off

Training – Testing	Precision	Recall	F1 score	TP identified	Accuracy*
ELM_owner	0.40	0.06	0.11	4/10	0.40
ELM_owner_noDup	0.80	0.13	0.22	8/10	0.80
ELM_followerTLDA	0.40	0.06	0.11	4/10	0.40
ELM_followerTLDA_noDup	0.40	0.06	0.11	4/10	0.40
ELM_followerFV	0.30	0.05	0.08	3/10	0.30
ELM_followerFV_noDup	0.36	0.06	0.11	4/10	0.40

*The accuracy is based on the true positive (TP) identified

Table 8. Results of the SVM – top 10 score cut off

Training – Testing	Precision	Recall	F1 score	TP identified	Accuracy*
SVM_owner	0.54	0.10	0.16	6/10	0.60
SVM_owner_noDup	0.70	0.10	0.19	7/10	0.70
SVM_followerTLDA	0.40	0.06	0.11	4/10	0.40
SVM_followerTLDA_noDup	0.40	0.06	0.11	4/10	0.40
SVM_followerFV	0.20	0.03	0.05	2/10	0.20
SVM_followerFV_noDup	0.25	0.05	0.08	3/10	0.30

*The accuracy is based on the true positive (TP) identified

4.4 Results of Using Top 30 Scores as Cut Off

Tables 9 and 10 show the results of using top 30 scores as cut off for the ELM and SVM respectively. From the tables, a similar trend but with higher accuracy can be observed when using the owner as the training dataset and when no-duplicate test data is used.

Table 9. Results of the ELM (based on the average of 10 runs, using 250 neurons) – top 30 score cut off

Training – Testing	Precision	Recall	F1 score	TP identified	Accuracy*
ELM_owner	0.50	0.24	0.32	15/30	0.50
ELM_owner_noDup	0.53	0.29	0.37	18/30	0.60
ELM_followerTLDA	0.39	0.21	0.27	13/30	0.43
ELM_followerTLDA_noDup	0.32	0.16	0.21	10/30	0.33
ELM_followerFV	0.23	0.11	0.15	7/30	0.23
ELM_followerFV_noDup	0.26	0.14	0.19	9/30	0.30

*The accuracy is based on the true positive (TP) identified

Table 10. Results of the SVM – top 30 score cut off

Training – Testing	Precision	Recall	F1 score	TP identified	Accuracy*
SVM_owner	0.47	0.22	0.30	14/30	0.47
SVM_owner_noDup	0.53	0.25	0.34	16/30	0.53
SVM_followerTLDA	0.43	0.21	0.28	13/30	0.30
SVM_followerTLDA_noDup	0.33	0.16	0.22	10/30	0.33
SVM_followerFV	0.25	0.13	0.17	8/30	0.27
SVM_followerFV_noDup	0.27	0.13	0.17	8/30	0.27

*The accuracy is based on the true positive (TP) identified

4.5 Comparing the ELM and SVM

The following two figures (Fig. 1 and Fig. 2) clearly show that the trend of F1 scores obtained for both the ELM and SVM is similar. As can be seen in the figures, the highest scores are found using the same configuration – owner as the training dataset

and no-duplicate as the testing dataset. In general, configurations based on owner as training datasets have yielded better results in predicting the target audience for both the ELM and SVM.

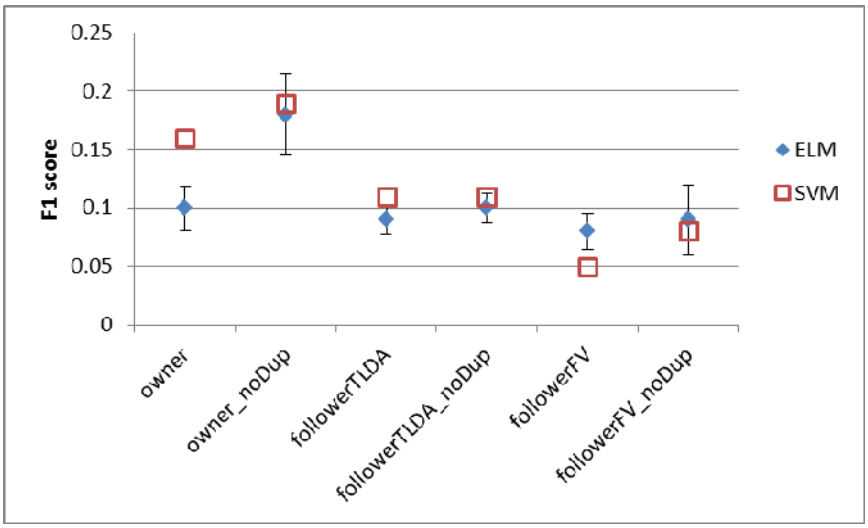

Fig. 1. F1 scores for the ELM and SVM based on top 10 score cut off. Error bars on the ELM indicate the 95% confidence intervals based on the Student T distribution of 10 runs.

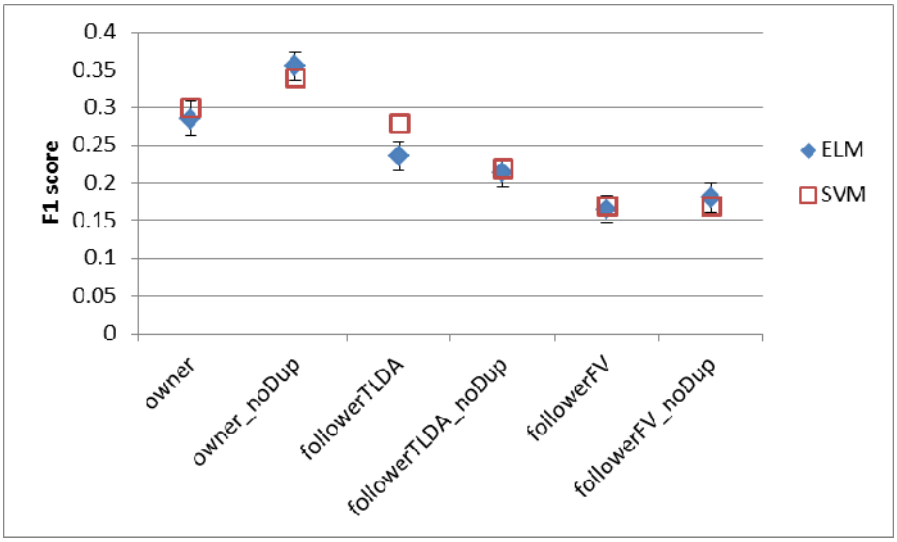

Fig. 2. F1 scores for the ELM and SVM based on top 30 score cut off. Error bars on the ELM indicate the 95% confidence intervals based on the Student T distribution of 10 runs.

5 Discussion

It is interesting to observe that, while traditionally a training dataset of the same source is often used for testing purposes, the F1 score results from top 10 and top 30 cut offs showed that using tweets from the account owner can yield better performances compared to tweets extracted from the list of followers when predicting the target audience. This finding is important as it eliminates the need to manually annotate the vast amount of tweets from the followers. Using tweets from the owner (which is well categorized within its domain) is more practical if it is to be adopted in a real-world application for target audience prediction.

The main objective of this study is to ascertain the effects of using training datasets built from either the owner or the list of followers for prediction. The results indicate that the types of training datasets have a clear impact on the outcome of the prediction process. Also, the preprocessing of the test dataset (i.e., having no duplication) is equally important in yielding better performances.

We have explored two approaches, namely Fuzzy match and Twitter LDA, for extracting representative features from the followers' tweets in this paper. The main reason of using these two approaches is because of their ability in enabling the annotation of followers' tweets through minimum manual efforts. Moreover, Fuzzy match and Twitter LDA (when used in conjunction with seed words extracted from the owner's tweets) have been shown to perform better than other methods in identifying the potential high-value social audience in our preliminary investigation.

It can be argued that the short list of seed words generated from the owner's tweets may not be able to form a feature vector that is representative of the tweets from the followers under the followerFV training dataset (using Fuzzy match). However, as the account chosen - "samsungsg" - is a technology and mobile company, its tweet contents tend to share specific terms such as products or events, which are essentially keywords that can be found in the tweets of the followers. In addition, we have analyzed both the training and testing datasets during the same period of time. It is therefore likely that the target audience who are interested in the content tweeted by the account owner will be tweeting similar terms or text. Having said that, this may not be the case for more generic accounts such as parent groups or current affairs, as the contents shared can be rather diverse and conceptual. As such, a more sophisticated feature generation approach based on domain-specific and common-sense knowledge may be required to enrich the bag of words [39] with new, more informative features.

While the results from both the ELM and SVM show similar trends, it is worthwhile to note that the computational time required by the ELM for training and testing is within the range of seconds. As this work focuses on the effect of using different training datasets for the prediction, however, we did not do a comprehensive comparison on the time spent in generating the model and the result between the two approaches.

We have compared the performances of the ELM and SVM mainly based on precision, recall and the F1 score. Considering the fact that companies and organizations are generally more interested in knowing which followers would more likely be interested in their products, we also used e scores and v scores generated from both the

ELM and SVM to identify the top 10 and 30 followers (see Tables 7, 8, 9 and 10). The ELM has performed slightly better than the SVM in this regard, as it succeeded in identifying more top followers by using tweets from the owner as the training dataset and no-duplicate data for testing.

6 Conclusion and Future Work

In this paper, we have used various training and testing datasets on both the ELM and SVM to predict and identify the target audience from a list of followers of a Twitter account owner "samsungsg". Our main purpose was to study the effect of using different training datasets to ascertain an approach for classifying the target audience with minimum annotation efforts.

From the results, we have observed that using the owner's tweets as the training dataset can better predict or classify the target audience than using the followers' tweets. In addition, it is essential to remove all the duplicates from the testing dataset, as this has shown to be able to improve the classification results. Equipping a company or organizer with the ability to predict the target audience enables the Twitter account owner to devise their marketing or engagement plan accordingly, in order to maximize the use of allocated budget and successfully reach out to potential customers in the crowded social media space.

Our future work will concentrate on enriching the features that can be identified from tweets, as the poor performance of Fuzzy match using the list of seed words extracted from the owner's tweets may be due to the limitation of bag of words. While Fuzzy match is able to identify terms without the need to have an exact match, it is not able to identify terms that it has not seen before. For example, Fuzzy match can identify "galaxy s iii" with "galaxy s3" but it is not capable of associating that "note 2" is also a related product. As such, it is essential to enrich the bag of words by incorporating online domain knowledge [40] (such as Wikipedia), integrating community-curated online databases [41] (such as Freebase) or combining entities from e-commerce sites (such as eBay) in order to form a more comprehensive view and thus improve on the prediction outcome.

References

1. 2013 Fortune 500 - UMass Dartmouth, http://www.umassd.edu/cmr/socialmediaresearch/2013fortune500/
2. How Many People Use Facebook, Pinterest, Twitter and 500 of the Top Social Media?, http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/#.U6kUVxAy1vA
3. Breslin, J.G., Passant, A., Vrandečić, D.: Social semantic web. In: Handbook of Semantic Web Technologies, pp. 467–506. Springer (2011)
4. Torres, D., Diaz, A., Skaf-Molli, H., Molli, P.: Semdrops: A social semantic tagging approach for emerging semantic data. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 340–347. IEEE (2011)

5. Yang, T., Lee, D., Yan, S.: 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)
6. Ramos, J.: Using tf-idf to determine word relevance in document queries. Presented at the Proceedings of the First Instructional Conference on Machine Learning (2003)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
8. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (1999)
9. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) Machine Learning: ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
10. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. Presented at the Proceedings of the Seventh International Conference on Information and Knowledge Management (1998)
11. Yang, Y., Liu, X.: A re-examination of text categorization methods. Presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1999)
12. Cambria, E., Mazzocco, T., Hussain, A.: Application of multi-dimensional scal-ing and artificial neural networks for biologically inspired opinion mining. Biol. Inspired Cogn. Archit. 4, 41–53 (2013)
13. Cambria, E., Huang, G.-B., Kasun, L.L.C., Zhou, H., Vong, C.-M., Lin, J., Yin, J., Cai, Z., Liu, Q., Li, K.: Extreme learning machines. IEEE Intell. Syst. 28, 30–59 (2013)
14. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011)
15. Mo, J., Kiang, M.Y., Zou, P., Li, Y.: A two-stage clustering approach for multi-region segmentation. Expert Syst. Appl. 37, 7120–7131 (2010)
16. Namvar, M., Khakabimamaghani, S., Gholamian, M.R.: An approach to opti-mised customer segmentation and profiling using RFM, LTV, and demographic features. Int. J. Electron. Cust. Relatsh. Manag. 5, 220–235 (2011)
17. 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)
18. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. 110, 5802–5805 (2013)
19. 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
20. Zhang, Y., Pennacchiotti, M.: Predicting purchase behaviors from social media. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1521–1532. International World Wide Web Conferences Steering Committee (2013)
21. Pennacchiotti, M., Popescu, A.-M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 430–438. ACM (2011)
22. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attrib-utes in twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, pp. 37–44. ACM (2010)

23. Ikeda, K., Hattori, G., Ono, C., Asoh, H., Higashino, T.: Twitter user profiling based on text and community mining for market analysis. Knowl. Based Syst. 51, 35–47 (2013)
24. Hong, L., Doumith, A.S., Davison, B.D.: 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, pp. 557–566. ACM (2013)
25. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the Fourth Workshop on Analytics for Noisy Un-structured Text Data, pp. 73–80. ACM (2010)
26. Konopnicki, D., Shmueli-Scheuer, M., Cohen, D., Sznajder, B., Herzig, J., Raviv, A., Zwerling, N., Roitman, H., Mass, Y.: A statistical approach to mining customers' conversational data from social media. IBM J. Res. Dev. 57, 14:1–14:13 (2013)
27. Using the Twitter Search API | Twitter Developers, https://dev.twitter.com/docs/using-search
28. Nakatani, S.: language-detection - Language Detection Library for Java - Google Project Hosting, http://code.google.com/p/language-detection/
29. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a max-imum entropy part-of-speech tagger. Presented at the 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)
30. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, pp. 985–990. IEEE (2004)
31. Huang, G.-B., Chen, L., Siew, C.-K.: Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans. on Neural Netw. 17, 879–892 (2006)
32. Huang, G.-B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for re-gression and multiclass classification. IEEE Trans. on Syst. Man Cybern. Part B Cybern. 42, 513–529 (2012)
33. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167 (1998)
34. Predictive Analytics, Data Mining, Self-service, Open source - RapidMiner, http://rapidminer.com/
35. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1015–1021. Springer, Heidelberg (2006)
36. Willett, P.: The Porter stemming algorithm: then and now. Program Electron. Libr. Inf. Syst. 40, 219–223 (2006)
37. Yang, M.-C., Rim, H.-C.: Identifying interesting Twitter contents using topical analysis. Expert Syst. Appl. 41, 4330–4336 (2014)
38. Kondrak, G., Marcu, D., Knight, K.: 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).
39. Harris, Z.S.: Distributional structure. Word (1954)
40. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. Presented at the IJCAI (2005)
41. Lo, S.L., Mei, S.D., Liew, V.: Use of Semantic Co-relation in Target Audience Profiling. Presented at the Fourth Global Congress on Intelligent Systems, GCIS (2013)