2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering

Identifying High Value Users in Twitter based on Text Mining Approaches

Yibing Yang, M. Omair Shafiq,
School of Information Technology,
Carleton University,
Ottawa, Ontario, Canada.
Emails: yibingyang@cmail.carleton.ca, omair.shafiq@carleton.ca

***Abstract*— Finding out new potential users for specific products are always the needs of the marketing department in industries. While Traditional ways like RFM model perform poorly in exploring new users. While the popularity of social media like Twitter and Facebook provides advertisers a new way to find, understand and target their users. In this paper, we propose a new method to find out and rank high-value target audience for a specific brand by utilizing machine learning and text mining approach. Overall tweets from 10 accounts in Twitter are collected to build the target and non-target dataset. In order to solve the data imbalance problem, five data resampling methods are assessed. Ensemble learning approaches include Bagging and Boosting algorithm are used to build the classifier. The results show that SMOTE outperforms other resampling method and AdaBoosting algorithm outperform other single classifier and Bagging model. We also find out the existence of marking accounts exists so that a threshold is set to filter these accounts which are not real users. We believe that our approach could be used in industry for identifying high-value users for online marketing purpose.**

***Index Terms*— Class Imbalance, Machine Learning, Text Mining, Twitter, User Classification, Social Media**

I. INTRODUCTION

The era of big data is coming, and this trend influences nearly every industry profoundly. It not only changes the way we create, transmit, store and process data radically but also change the way we understand and analyze the data. The industries which people can feel the revolutions directly are areas like media, communication, and information because they are closely related to our daily life. One of the biggest change is the emergence of social network which is a sign of the coming of web 2.0. It changes the way how information transmitted and offer everyone who is able to get access to the internet the opportunity to express their thoughts and ideas. Compared with the traditional media like TV stations, radio stations, newspapers, and magazines which keep the dominating position in spreading information to the audience, social media is a user-generated-content platform where people express their ideas freely about nearly anything they want with low cost instead of receiving information passively. With the features of inclusiveness and de-centration, it is the first time we have the opportunity to get access to the data from individuals directly. It also brings more commercial opportunity for companies which are eager to know more about their consumers.

Traditionally, if advertisers or companies want to advertise, they tend to go for media like TV stations or newspapers while now they tend to advertise online. This is first because the social network is becoming more and more influential these days and the traditional media is losing their dominating positions. Another reason is that for advertisers it is hard to measure the advertising effect from traditional media because they cannot choose the target audience and they cannot build the relationship between people who saw the advertisement and people who went to shopping either. In order to get to know their consumers, market research companies would be hired by them to do research based on sampling in terms of the demographic information. Compared with this time consuming and trouble consuming procedure, online marketing based on social media gets a variety of benefits. Firstly, it helps companies to save a lot of money which they spent on traditional media that broadcast to everyone watching TV or reading the newspapers but not only the audience they want to target. Secondly, advertisers and companies can get in touch with the users via social media by which they could acquire fruitful information and conduct personalized recommendation in terms of their interests to achieve a win-win situation.

After comparing several existing approaches, we propose a new system which could be regarded as a pipeline to identify high-value audience with the least human manually efforts. We choose a particular official account of a brand in our study based on previous work [12] while this approach could also be applied to target audience detection for other brands.

II. RELATED WORK

Companies want to make profits by finding the right target users which requires user classification. Traditionally, there are two kinds of ways to achieve this purpose. One possible way is to focus on data from user behaviors which are mostly from CRM system or transaction data. When it comes to the CRM system, RFM (recency, frequency, and monetary) model and demographic information are the most commonly used variables for consumer classification [1]. In order to take advantage of the demographic information from the social network, [2] infers users' demographic attributes from limited information in the social network. They found out that user attributes could be inferred with high accuracy even when only 20% information about the users was provided.

In [3], authors have done an exploratory on Twitter user attribute detection. The authors tried to deduce latent attributes

2324-9013/19/$31.00 ©2019 IEEE
DOI 10.1109/TrustCom/BigDataSE.2019.00090

of users, network structure, and communication behavior and made some cross analysis. The classification models were employed including SocioLinguistic-feature Models, Ngram-feature Models and Stacked Model based on SVM algorithm to solve the classification tasks with different combinations of the attributes. In [4], an algorithm was proposed to improve the performance of user classification. The results showed that the system including the clustering features outperformed the N-gram+Token system. Meanwhile, all the differences between using all features and N-gram+Token features are significant.

In [5], authors found out correlations between users' social media profiles and their e-commerce behaviors. The results showed that the performance of demographics features was not as good as other features and Facebook Categories features achieved the best accuracy while n-gram got a reasonable result. In [6], authors present a study of Twitter user classification based on Gradient Boosted Decision Trees in terms of three different tasks about political affiliation detection, ethnicity identification and detecting affinity.

The study [7] could be regarded as a subsequent research of [6] which continuously focus on user classification on Twitter. The novel part of this research is that based on the features which had been discussed in the previous work, a graph-based updating component which integrated social network information experimented with the assumption that social connections can help to correct the errors of the results of machine learning classifier by inverting the classification label to the correct ones. [8] purposes that most of the previous work focused on the network or community structure in Twitter while few works have been done with the textual content of posts on Twitter. In [9], authors aimed at automatically establishing who is participating in information production or conversation around events by building an automatic classifier for user types on Twitter. Specifically, initial steps were taken towards building an automatic classifier (kNN with k = 10) to categorize user into core types. In [10], a series of text mining and machine learning approaches were used to predict and identify target audience from a list of followers of a specific account owner on Twitter. In [11], the authors continued their research in [10]. They used Twitter API to get the data (Samsung Singapore or "samsungsg" dataset) for analysis. While compared with their former work that fuzzy match and Twitter LDA are applied to extract features from tweets as well as ELM and SVM are machining learning algorithms to building models and assess the performance. In a later study [12], the authors kept focusing on identifying target audience in the social network based on an index system which was built by utilizing Fuzzy Match, Twitter LDA and SVM. The results showed that the HVSA index is a better indicator than individual scores from the various methods.

While in [13], the authors give an overview of the approaches that were used in [10], [11] and [12]. It explained the theory behind text mining approaches including fuzzy keyword match method, Twitter LDA method and Machine learning like Support Vector Machine and the way how to integrate them into separating followers and describing a group of high-value social audience members. In [14], the authors were aiming at improving the performance of their existing algorithms. In this research unsupervised (Twitter LDA) and supervised (SVM ensembles) learning methods were applied to automatically classify and identify a target audience from a list of followers of a Twitter account. In [15], the authors integrated methods they used in previous work to achieve a better result. In order to find out high-value social audiences in Twitter, LDA, Fuzzy Match, SVM ensembles, and HVSA ranking which is based on the previous three methods were applied. In this research simple average schema and linear regression schema was experimented to evaluate the results of the models.

In [16], a demographic estimation algorithm was presented to predict the profile information of Twitter users. This purpose was achieved by employing a hybrid text-based and community-based method for the demographic estimation of Twitter users, where these demographics are estimated by tracking the tweet history and clustering of followers or followees. The result showed that this new approach could outperform the text-based one and even lead to a desirable accuracy for inactive users. In [17], approaches to automatically categorize users' professions and personality related attributes were explored. The results showed that for both personality and professional classification, linguistic style features are most useful feature while social-semantic features are very useful for predicting personality related attributes did not achieve desirable accuracy for professional areas.

In [18], the researchers tried to find out users' interests by examining the entities they mention in their Tweets. With a knowledge base built based on Wikipedia, they managed to disambiguate and extract the entities in the Tweets. After that, a "topic profile" was built in order to characterize users' topics of interest in a way of discerning which categories appear frequently and cover the entities. In [19], authors aimed at understanding the conversation which was expressed in social media. A scalable user-profiling solution was utilized and implemented using the Apache Hadoop framework.

In [20], authors focused on classifying political orientation of users on Twitter. They proposed that the accuracies of previous accuracies have been overoptimistic due to the way in which validation datasets have been collected with an approximately 30% higher than expected. In [21] authors integrated different methods within a single framework to facilitate the attribution and differentiate Online Social Networks members. They also examined how peer effects influenced the expressions of users.

III. Methodology and Proposed Solution

Compared with the traditional way by using features to predict demographic features to segment the users, mining the semantic and network structure on social media based on supervised machine learning algorithm is a better way to understand the users' opinion and needs. While there are still some questions about the prevailing method.

1) Lack of open datasets: We have reviewed a series of papers about user classification on Twitter. However, different datasets have been used in different studies. Some of them are collected from Twitter directories, some of them are crawled by the researchers themselves while some are an integration of different datasets. So that the question here is that it is hard to evaluate the results of different methods because they all get different data sources.

2) Data imbalance problem: It means that the instances of different categories in training dataset account for significantly different portion. The potential consequences of data imbalance are that making biased predictions and getting misleading accuracy. While this problem has not been discussed explicitly in previous works.

3) Lack of definition of high-value users: In previous works, there is a gap between detecting highly related tweets and high-value users. There is fewer explanations about how to use the highly related tweets to target the high-value users. In [12], authors present a possible way by calculating the percentage of highly related tweets accounting for the number of all tweets posted by a user. While this way may not work well in users who have diverse interests. No researchers so far explain how they label the test data.

4) High dimension of text data: A framework of our plan is shown in Figure 1. The first phase is data collection part. The names of target brand and non-target brands would be input by us. The name of target brand is chosen by us depend on which brand we want to look deep into it. While non-target should be other brands which could be explicitly identified as a totally different area and topic by humans. Then for both target and non-target brands, we are going to collect all the tweets they posted and all the tweets from their active users.

Then two kinds of datasets would be built: target dataset and non-target dataset. We could utilize the two sorts of tweets we get as positive training data and negative training data with a significant difference. After that, we are going to build a scoring system to evaluate the results of the target audience detecting. Fuzzy match, Twitter LDA, and SVM are chosen as three text mining approaches for assessing the similarity of the tweets between training dataset and test dataset. The benefit of doing this is that we could rank the audience in test dataset based on the combination of three kinds of scores which could help us to identify the potential users from high to low ranking.

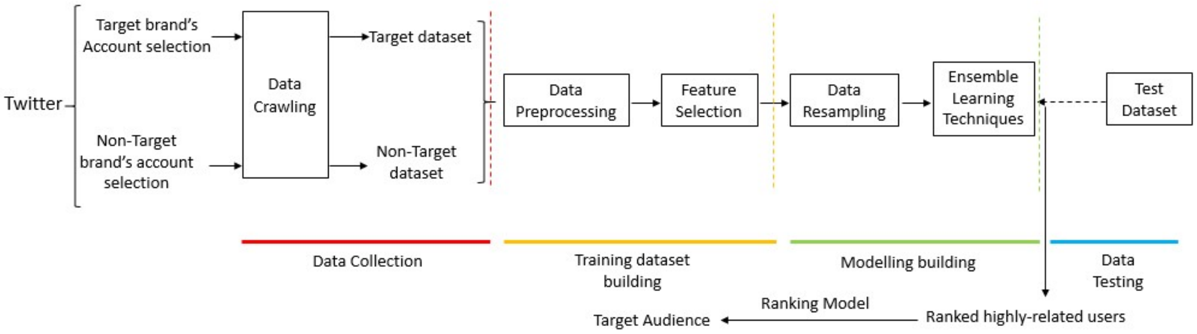

Figure 1. High Value Target Audience Detection Framework

A. Data collection

1.1 Target Twitter accounts selection

In order to test the validity of our proposed approach, we are supposed to find out a group of Twitter accounts for our research. The target Twitter account we select is 'SamsungSG' which is the official source for Samsung products and news in Singapore. This official account has been chosen as the target account in a series of previous studies. In order to compare the result of our study with others work, we choose it as well in our paper while the data from this account is collected by our own which is going to be discussed later. The details of selecting accounts could be reviewed in work by Lo et al.[22].

In previous work, only 200 tweets from a Twitter account were used because of temporal problem. While in our research, we assume that the topic for a specific Twitter account would not change dramatically so that we tend to use more tweets than previous researches.

Table 1. Collected Datasets

Account type	**Account name**	**Field**	**Tweets collected**
Target	SamsungSG	Electronic	550
Non-Target	MOEsg	education	819
Non-Target	JoannePeh	celebrity/daily musing	534
Non-Target	Camemberu	food/travelling	612
Non-Target	premierleague	football	811
Non-Target	sgbroadcast	news	811
Non-Target	mtvasia	music	825
Non-Target	SGAG_SG	social media	801
Non-Target	ilovedealssg	shopping	780
Non-Target	kiasuparents	Parenting/ daily musing	823

1.2 Data crawling method

We have used our own data scraping program based on Selenium, Request, re, BeautifulSoup packages in Python to collect the data we need. For each account, we collect their tweets, posted time and posted date.

1.3 Topic Modelling

After solving the problem to collect and record all these data, a remaining question is that it is still challenging for us to organize, search, and understand the content of them and get access the knowledge we want. In order to examine the diversity of the content posted by the account owners, we purpose to utilize the topic model to check the model distribution of the tweets from different accounts. Topic Modelling is a popular approach in text mining field which is introduced to analyzing vast quantities of information. It is always used to help to comprehend and to extract the potential meaning of a series of documents. It is a kind of probabilistic model provide researchers a way to discover the hidden topics, annotate documents according to topics and utilize the annotations to summarize the big volume of documents. The topic model is based on bag-of-words model and the general idea of it is to transfer the input data based on bag-of-word distribution into a lower dimensional bag-of-topic format.

The earliest implementation of the topic model is latent semantic analysis (LSA) [23] based on matrix. After that probabilistic latent semantic analysis (pLSA) [24], a model evolved from the previous model by adding the probabilistic model, was proposed by Homas Hofmann in 1999. While pLSA was considered incomplete in that "it provides no probabilistic model at the level of documents.", the arising of Latent Dirichlet Allocation (LDA) [25] solve this problem. Using this, we acquired the results of the topic representations of all the documents and word distributions of all the topics when a stable state is achieved.

B. Data Preprocessing

Data preprocessing is an essential part of text mining especially for tweets which directly determine the results of the following procedures. Due to short length of tweets, high volume of noises, usage of the irregular word, symbols, abbreviations, it is harder to process than text data like news and blogs which get a longer paragraph and regular wording. We apply a variety of methods for data preprocessing part.

2.1 Data Cleaning

First, we implement the data cleaning on our tweets. For each tweet in all Twitter accounts aforementioned, first we drop the duplicated tweets and strip the blank on the left and right of the tweet; After we tokenize the words and lowercase them; Then the HTML symbols are removed, so as the mention (@), hashtags (#) symbols and URLs embedded in tweets; In addition, numbers and punctuations appeared in tweets are dropped; Words which are shorter than one letter are also removed in this section.

2.2 Tokenization

Tokenization is a process in which each tweet in the dataset will be separated into units which are called tokens. In our study, the units are words in tweets and each token is a word. In natural language processing, tokenization is always used as a preprocessing step before further analyzing.

2.3 Data Normalization

In this step, we turn to the data normalization part. First we standardize the words in tweets (e.g. work like 'goooood' would be standardized to 'good'); Then we filter all the stop words appeared in tweets; At last, we lemmatize the words (e.g. different form of words would be lemmatized to the same form like 'gets' and 'got' would all be changed to 'get').

C. Feature selection

After we get a clean dataset from previous sections, we could go further to the feature selection part. This part also directly determines the input of the following algorithm. Therefore, we propose different ways to select proposed features for the following procedures:

3.1 POS tag

POS is the abbreviation of part of speech which is also a commonly used way to tag text data. In traditional, the part of speech of word includes noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection and article. In text mining tasks words with POS tags of nouns are always used to find out the potential topics among text because nouns from different topics tend to be more different compared with other parts of speech. While adjective tags tend to be used in sentiment detection because they are more related to emotional expression. We are going to discover how different selection of POS would affect the result of our algorithm.

3.2 Vector space model

Vector space model is an algebraic model which could represent documents and words in them into vector space. Traditionally in a vector space model, the rows are different documents and the columns (dimensions) are different words. If a word in the dimensions occurs in a certain document, the value of it will be non-zero. In text mining, the values could be filled with term frequency or TF-IDF which will be discussed in the next section. In this way, a series of documents could be represented in vectors that could be calculated.

Table 2. Vector Space Model

	word1	word2	…	wordN
Tweet1	Value1.1	Value1.2	…	Value1.N
Tweet2	Value2.1	Value2.2	…	Value2.N
…	…	…	…	…
TweetN	ValueN.1	ValueN.2	…	ValueN.N

While this vector space is usually quite sparse when dealing with text data because most of the values in the table tend to be zero. In order to solve this problem chi-square test would be used to reduce the dimensions.

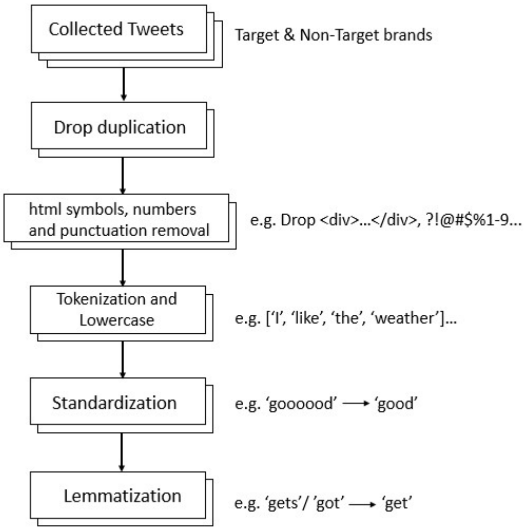
Collected Tweets
Target & Non-Target brands
Drop duplication
html symbols, numbers and punctuation removal
e.g. Drop <div>…</div>, ?!@#$%1-9…
Tokenization and Lowercase
e.g. ['I', 'like', 'the', 'weather']…
Standardization
e.g. 'goooood' ⟶ 'good'
Lemmatization
e.g. 'gets'/ 'got' ⟶ 'get'

Figure 2. Data Processing procedure

3.3 TF-IDF

TF-IDF is the abbreviation of term frequency–inverse document frequency, which is one of the most popular termweighting schemes. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.[26] Term frequency represents the frequency of a word appearing in a document while the inverse document frequency reflects the uniqueness of the word among different documents.

As we used TF in our earlier work [32], we assign TF to the words appears in tweets with *w* is the word, *t* is the tweet contains *w*, and *f* is the frequency. So that the term frequency could be calculated by:

$$tf(w,t) = \frac{f_{w,t}}{number\ of\ words\ in\ t} \qquad (1)$$

For the inverse document frequency, we assign *T* as the collection of tweets while *N* is the total number of tweets.

$$idf(w,T) = \ log \frac{N}{|t\in T, w\in d|} \qquad (2)$$

Then TF-IDF is: $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$ *(3)*

3.4 Chi-square test

There are several ways for feature selection in text classification tasks like Consistency, Information Gain, Document Frequency, CFs and so on. While Chi-square Test has been proved to perform better compared with other approaches [26]. The main idea of Chi-square Test is to check if the differentiation between expected frequencies and the observed frequencies is significant or not. Traditionally we could set null hypothesis as two variables are independent of each other. Then we could calculate the differentiation with the assumption that they are independent. If the differentiation is small enough, we could consider that the differentiation is caused by sample error. Otherwise, if the differentiation is bigger than a specific threshold, we would refuse the null hypothesis and accept H1 that the variables are dependent.

D. Data imbalance problem in Training Dataset Building

In this part we state the data imbalance problem which could impact the performance of the classification algorithm and proposed three different ways for data resampling.

4.1 The universal data imbalance problem in machine learning

In our project, one major concern in building the training data is the data imbalance problem. Data imbalance is not always a problem for all the machine learning tasks but it probably impacts the accuracy of some specific tasks especially when it related to detecting a particular pattern from big amount of dataset. To be more specific, we are going to finding out target users for a specific brand, so the positive training data is the tweets highly related to this particular brand (in our study, we are using the tweets of SamsungSG). While the negative training data is the tweets from other topics other than tweets about mobile phone and electronic products. So that it is obvious that the volume of positive training data (tweets about samsung) is limited while the volume of negative training data is much more available than the positive counterpart.

4.2 Data Resampling Approaches

To solve the data imbalanced problem discussed in the previous section, several resampling approaches are used.

Random over sampling (ROS) our training dataset with the non-target class and target class which the size of target dataset is always far less than the size of the non-target dataset. The theory behind this approach is to increase the size of the minority class to the one of the majority class based random sampling. Then the new over-sampled dataset would be used as the new dataset.

Synthetic Minority Over-Sampling Technique (SMOTE) [27] is one of the over-sampling approaches to solve the data imbalance problem. In this method, the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement. The synthetic examples cause the classifier to create larger and less specific decision regions.

Adaptive Synthetic (ADASYN) [28] is another Over-sampling approach. The theory behind ADASYN is to assign different weights on different examples in minority class. To be more specific, more synthetic data would be generated for the examples which are hard to be classified while less would be generated for the ones which are easier to learn. By using ADASYN, the re-balanced dataset would be reduced the bias and shift the classification decision boundary for the ones which are hard to be classified.

Random under sampling (RUS) gets our training dataset with the non-target class and target class which the size of target dataset is always far less than the size of the non-target dataset. The main idea of this approach is to reduce the size of the majority class to the one of the minority class based random sampling. Then the new under-sampled dataset would be used as the new dataset.

Under-sampling Cluster Centroid is an Under-sampling method which generates centroids based on clustering approach. By applying Under-sampling Cluster Centroid, the majority class would be replaced by a cluster of examples which are the cluster centroid of the K-Means algorithm [29]. This algorithm keeps majority samples by fitting the K-Means algorithm with clusters to the majority class and using the coordinates of N cluster centroids as the new majority samples.

E. Machine learning algorithms and ensemble learning

Ensemble learning is a popular approach to machine learning these years. In ensemble learning, we look at multiple classifiers and combine the output of the multiple classifiers in order to get better classification accuracy. The classifier outputs are independent of each other and make errors in an independent manner, it is possible that by combining the outputs of several classifiers. We get a resulting classifier which is better than any of the constituent classifiers.

Compared with trying to improve the accuracy of one model, ensemble methods take a series of models into account and take the average of those models to make the final decision. In this study, we propose Bootstrap aggregating [30] and Ada-Boosting [31] to solve our text classification problem.

When solving machine learning tasks, some techniques are sensitive to the variation in the training dataset. While Bootstrap aggregating [30] is a way to reduces variance and helps to avoid over-fitting. The main idea of Bootstrap aggregating method is to produce several new datasets by sampling from the original tweet dataset. After producing the new datasets, different classifiers will be built based on these different dataset and diverse results would be achieved. In the final part, the results of classification are prepared by the combination of the results of different classifiers in terms of a major vote scheme, as depicted in Figure 4.

In our study we choose Adaboosting [31] algorithm for ensemble learning. The main idea of Adaboosting is to initialize the weights distribution in the first round with equal score. Then in the first round, the original dataset is sampled based on initial weight distribution and a classifier is built on it. After that, the result will be tested on the original dataset. While the initial weights of features would be updated with the standard that the weights of the ones which were wrongly predicted will be raised and the rest which was correctly classified will be decreased. Then a new round would be started based on the resampling dataset. The number of rounds should be decided at the beginning and the final results are the linear combination of the different classifiers.

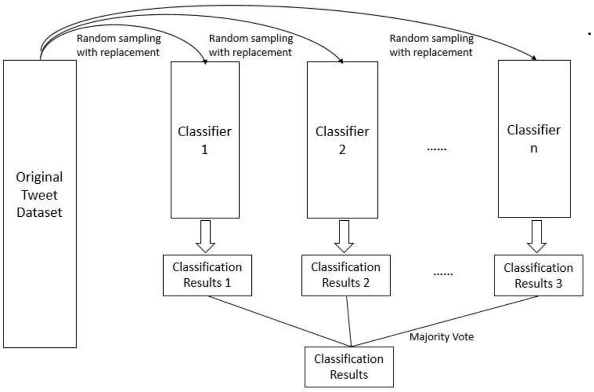

Figure 3. Bootstrap aggregating process

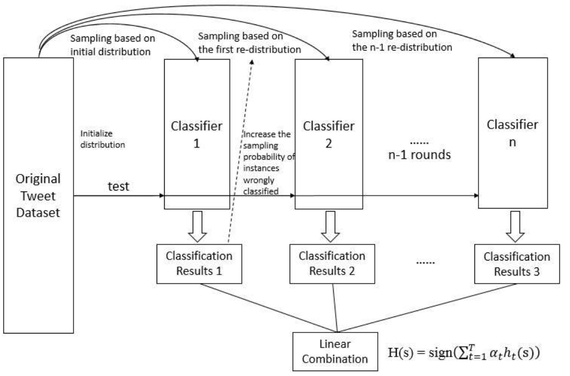
Sampling based on initial distribution
Sampling based on the first re-distribution
Sampling based on the n-1 re-distribution
Original Tweet Dataset
Initialize distribution
test
Classifier 1
Increase the sampling probability of instances wrongly classified
Classifier 2
......
n-1 rounds
Classifier n
Classification Results 1
Classification Results 2
......
Classification Results 3
Linear Combination
$H(s) = sign(\sum_{t=1}^{T} \alpha_t h_t(s))$

Figure 4. Adaboosting process

Our work, in contrast with the previous studies, only use the content feature which is the information extracted from the tweets posted by users. By doing this, we could reduce the requirement of graph computing resource when adding network features. Also, the low quality of demographic information and irrelevant behavior information are not considered in our study because we only want to focus on what the users are interested in from the text they posted on their own will. We discuss which content feature is most effective to identify high-value users.

In addition, we focus on the feature selection part to achieve a better result. We also test another ensemble learning approach, Boosting, which have not been tested in previous work. We want to compare the performance of Adaboosting, which is one method of Boosting algorithm, with other three single classifiers and another ensemble learning model, bagging, which has been assessed in previous work.

IV. EVALUATION

In this section, we discussed the evaluation of the previous methodology we proposed. First, we introduce the Evaluation Methods we use to measure our model. Then we assess whether our target and non-target datasets are different from each other or not by using term frequency distribution and topic modeling. After that, we explore how different features, different k values for chi-square test and different resampling approaches influence the classification. In the last section, we discuss how to use our model to detect high value of the target brand and the way we use to filter the marketing accounts.

4.1 Dataset Discussion

We have discussed how we build the training dataset we need from Twitter in previous sections. Overall the target brand dataset consists of 550 tweets while the non-target brand dataset contains 6816 tweets which are from nine different brands' account. For both the target brand dataset and non-target brand dataset, we run the data preprocessing procedure discussed in Section 4. As a result, 543 tweets and 5762 tweets remain respectively for the two accounts type.

Table 3. Target and Non-target datasets

Account type	Number of original Tweets	Number of Tweets after preprocessing	Label
Target	550	543	Positive
Non-Target	6816	5207	Negative

After the preprocessing, we want to ensure that the topics from target dataset and the non-target dataset are different from each other. For the purpose of examining, we use term frequency and topic model to analyze our datasets. First, we checked the term frequency distribution of the target brand and non-target brand. Before we calculate the most frequent word, we use POS filter to acquire only the nouns in the dataset because nouns could be the most representative words among different topics while other POS sometimes tend to be used repetitively in different topics and also bring some noises. We observed based on the result that the top words in target brands are all highly related to electronic devices and technology, e.g., 'galaxy', 'samsung', 'camera', 'app', and 'device'. while the top words in the non-target dataset are quite generalized like 'student', 'time', 'year', 'school', and 'world'.

As it shown in 0, the first column is the names of the brands for target brand and non-target brands. While he second column is the topic covered by each brand and the third column is the confidence of them. It is clear that each covers a variety of topics and the topics of the target brand are different from the ones in the non-target brand.

Table 4. Topic distribution of target and non-target accounts

Brand	Topic	Confidence

Samsung	Arts / Culture / Entertainment	87
	Technology	70
	Product Launches	45
	Software & IT Services	69
	Media & Publishing	42
	Product Launches	59
MOEsg	Education	97
	Arts / Culture / Entertainment	7
JoannePeh	Human Interest	95
	Media & Publishing	7
	Arts / Culture / Entertainment	7
Camemberu	Hospitality Recreation	93
premierleague	Sports	93
sgbroadcast	Government / Politics	94
mtvasia	Arts / Culture / Entertainment	100
	Media & Publishing	60
SGAG_SG	Arts / Culture / Entertainment	67
ilovedealssg	Hospitality Recreation	87
	Arts / Culture / Entertainment	77
kiasuparents	Arts / Culture / Entertainment	87
	Media & Publishing	78
	Children / Youth Issues	44

4.2 Different features

In this section, we assess the impact of using different features. Four different features are tested overall: Using nouns in tweets with Tf-Idf score, using nouns in tweets with term frequency score, using all the words with the Tf-Idf score, and using all the words with Tf-Idf score. The results are shown in Fig below. The other parameters we applied are as following: k value = 5000, resampling method: Smote, ensemble learning basic learner: Logistic Regression, number of classifiers: 7.

From the F measure, the results of using Tf-Idf outperform the ones applying only TF. In terms of part of speech, using noun only slightly outperforms using all the words. The best result appears when applying Noun_TFIDF feature on Boosting. Similarly, Bagging also gets its best result when using Noun_TFIDF. While for Logistic Regression, using all the words with TFIDF get the best the score which is even higher than the ensemble learning models. Interestingly, it shows an opposite trend when applying noun features on SVM. Using only noun leads to worse results compared with applying all the words features. As for Naïve Bayes, it gets a better result using TF no matter when applying noun features or all the word features. We could see that different algorithms have different feature preference.

4.3 High-value user detection

In this section, we discuss the way how we define a high-value user from others by running our model on the data of the followers of our target brand. The parameters of the model we choose in this part are as follows: k value = 5000, feature = Noun_Tfidf, Algorithm: Boosting, ensemble learning basic learner: Logistic Regression, number of classifiers: 7. The users we choose are 200 followers of target brand 'samsungsg'. After running data preprocessing on their tweets and filter the users who have at least more than 10 tweets in total, we get 139 users waiting for be ranked. For each user, we define a ratio to measure the likelihood of being a high value.

The number of target tweets of a user is the number of tweets which are predicted by the model as positive which the number of total tweets is the number of tweets a specific user has.

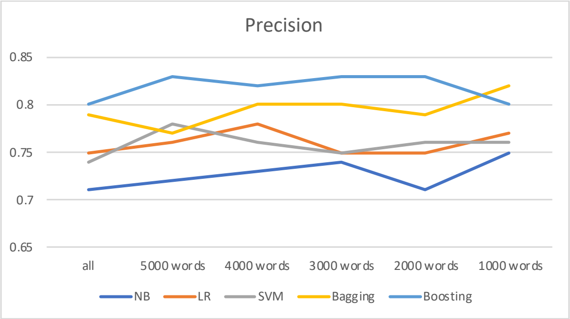

Figure 5. Precision of different k value in Chi-square test

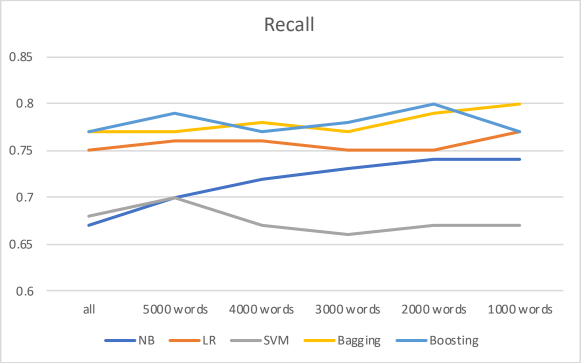

Figure 6. Recall of different k value in Chi-square test

4.4 Comparison with previous work

In this section, we would like to see the results of the 10-fold cross validation of our proposed solution (LR with bagging and LR with boosting) and the approach proposed in [14].

Table 5. Results of 10-fold cross validation

method	**recall**	**precision**	**F measure**
LR with bagging	0.98	0.97	0.98
LR with boosting	0.99	0.98	0.99
SVM with bootstrapping sampling	1.00	0.98	0.99
SVM with 10 random sampling, majority vote	0.31	0.46	0.37
SVM with bagging	0.69	0.97	0.80
SVM with stacking	0.96	0.90	0.93

V. CONCLUSION

In this study, we proposed a system which could be utilized to identify the high-value audience of a specific brand on Twitter. This approach has the potential to meet the needs of companies to target their audience with low cost by using only the text information in their Twitter account. In previous work, we have seen that in order to solve users or content classification problem, features including demographics, content, behavior, and network information have been used to optimize the results. While in this study, we only focus on content information which are tweets posted publicly by users on their Twitter accounts because the content is directly related to the interest of users. We built training dataset by analyzing the preferences of the users of our target brand, samsungsg, based on the topic model to find out the fields of the non-target brands which are quite different from the topics covered in target brands. We also discuss the details of feature selection procedures and evaluate the results when applying different parameter in Chi-square test. We also explicitly discuss the data imbalance problem in target audience identifying and assess five data rebalance methods on our dataset and we find that Smote outperforms other resampling approaches in our study. Apart from this, we compare three basic classifiers and two ensemble learning algorithm. The results show that boosting method outperforms other algorithms in most cases. For the classification part, we evaluate two ensemble learning algorithms which achieve desirable results.

REFERENCES

[1] Miglautsch J R. Thoughts on RFM scoring, journal of Database Marketing & Customer Strategy Mgmt, 2000, 8(1):67-72.

[2] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In ACM international conference on Web search and data mining, pages 251–260.ACM, 2010.

[3] Delip Rao, David Yarowsky, Abhishek Shreevats, Manaswi Gupta. Classifying latent user attributes in twitter. In intl workshop on search & mining user-generated contents, pages 37–44. ACM, 2010.

[4] Shane Bergsma, et. al., Broadly improving user classification via communication-based name and location clustering on twitter. In the conference of the North American Computational Linguistics: Human Language Technologies, 2013.

[5] Yongzheng Zhang and Marco Pennacchiotti. Predicting purchase behaviors from social media. In proceedings of WWW 2013, pages1521–1532.

[6] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. Icwsm, 11(1):281–288, 2011.

[7] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In ACM SIGKDD intl' conference on Knowledge discovery and data mining, pages 430–438. ACM, 2011.

[8] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. ICWSM, 10(1):16, 2010.2

[9] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. Unfolding the event landscape on twitter: classification and exploration of user categories. In ACM conference on Computer Supported Cooperative Work, pages 241–244, 2012.

[10] Lo S L, Cornforth D, Chiong R. Effects of Training Datasets on Both the Extreme Learning Machine and Support Vector Machine for Target Audience Identification on Twitter, in ELM-2014 Volume 1. Springer, 2015:417-434.

[11] Siaw Ling Lo, David Cornforth, Raymond Chiong. Identifying the high-value social audience from twitter through text-mining methods. In 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Vol 1, pages 325–339.Springer, 2015.

[12] Siaw Ling Lo, David Cornforth, Raymond Chiong. Use of a high value social audience index for target audience identification on twitter. In ACALCI, pp 323–336. Springer, 2015.

[13] Priyanka B Dastanwala and Vibha Patel. A review on social audience identification on twitter using text mining methods. In Wireless Communications, Signal Processing and Networking (WiSPNET), pages 1917–1920. IEEE, 2016.

[14] Lo S L, Chiong R, Cornforth D. Using support vector machine ensembles for target audience classification on Twitter[J]. Plos One, 2015, 10(4):e0122855.

[15] Siaw Ling Lo, Raymond Chiong, and David Cornforth. Ranking of high-value social audiences on twitter. Decision Support Systems, 85:34–48, 2016.

[16] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Systems, 51:35–47, 2013.

[17] Claudia Wagner, Sitaram Asur, and Joshua Hailpern. Religious politicians and creative photographers: Automatic user categorization in twitter. In Social Computing (SocialCom), 2013 International Conference on, pages 303–310. IEEE, 2013.

[18] Matthew Michelson and Sofus A Macskassy. Discovering users' topics of interest on twitter: a first look. In 4th workshop on Analytics for noisy unstructured data, pages 73–80. ACM, 2010.

[19] Konopnicki D, Shmueli-Scheuer M, Cohen D, et al. A statistical approach to mining customers' conversational data from social media. IBM Journal of Research & Development, 2013, 57(3/4):14:1-14:13.

[20] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In ICWSM, 2013.

[21] Muhammad Al-Qurishi, et. al., User profiling for big social media data using standing ovation model. In Multimedia Tools and Applications, pages 1–23, 2017.

[22] Lo SL, Chiong R, Cornforth D. Using support vector machine ensembles for target audience classification on Twitter. PLoS ONE. 2015; 10(4).

[23] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

[24] Hofmann, Thomas. "Probabilistic latent semantic analysis." Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999.

[25] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[26] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive datasets. Cambridge university press, 2014.

[27] Chawla, Nitesh, et al. "SMOTE: synthetic minority over-sampling technique." Journal of AI research 16 (2002): 321-357.

[28] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." In IJCNN 2008.

[29] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24.7 (2002): 881-892.

[30] Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.

[31] Freund, Yoav, Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of comp. and sys. sci. 55.1 (1997): 119-139.

[32] Yibing Yang, M. Omair Shafiq, "Large scale and parallel sentiment analysis based on Label Propagation in Twitter Data", in 12th IEEE Intl Conf On Big Data Science And Engineering (IEEE BigDataSE 2018), 1-3 Aug 2018, New York, USA.