Evaluating Classification Schemes for Second Screen Interactions

Partha Mukherjee
College of Information Science and Technology
Pennsylvania State University
University Park, PA, USA
pom5109@ist.psu.edu

Bernard J. Jansen
College of Information Science and Technology
Pennsylvania State University
University Park, PA, USA
jjansen@acm.org

***Abstract*— We analyze the performance of classification schemes on information collected from social conversation posted in Twitter among audiences of a popular US based TV show. In this research, we consider entropy as a measure of information exchange in a group conversation that is related to social, temporal, and second screen device features. The group conversations are identified by hashtags present in tweets where the number of members in the group interacting is at least two. We apply different classification schemes to more than 4,900 groups identified from 318,000 tweets. The result shows that 5-nn algorithm outperformed the other procedures in terms of misclassification error to identify the informed groups.**

Keywords— entropy; informed groups; misclassification; PCA; second screen; social TV

I. Introduction

With the advent of Internet technology and the emergence of online social networks, the social possibility of TV has greatly expanded, as the merging of technologies now allow a number of social activities and interactions concerning TV content via social networks (e.g., Facebook, Twitter, Weibo etc.).

The viewers do exchange information related to the TV show via second screen in terms of posting of tweets [1]. The exchange of information can happen live during show time or when the show is not transmitted live. The information in the social interaction may relate to different aspects of the show content (e.g., the actors, directors, costumes, characters, themes, etc.) or aspects of the products advertised (e.g., brand, sale, customer preference etc.). These attributes surrounding to a TV show in the discussion are identified by means of the hashtags (#) before a relevant keyword or phrase in the tweets to categorize the tweets as part of the popular topics marked by those keywords or phrases. Audiences reciprocating the tweet mentioning a particular hashtag to the commentary form a group engaing in a conversation. The group size is determined by the number of members interact within that group. The greater the interaction within a group, the more informed group it will be. Highly informed groups promote diffusion of more information about the aspects concerning a TV show.

In this research, we investigate the degree of information exchange in group interactions regarding a TV show using second screens. The investigation involves the classification of the groups into *informed categories* depending on the quantification of information interchange within the groups. The quantification of information is expressed in terms of social, temporal, and second screen device features of the conversation.

This research is important, as degree of second screen interaction will eventually determine the predominant group conversations with high information content. This insight will lead to the emergence of popular topics, which in turn can facilitate the personalization of TV content and advertising. The channel owners and advertisers need to correctly identify the paramount groups that corroborate the show or ad-related information with higher probabilities in order to prioritize their personalization efforts. Therefore, the findings assist both the broadcasters and advertisers to classify the leading interacting groups with minimum error that will help them to formulate new strategies for TV airing, launching product ads to engage more viewers, promote sales, and earn revenues.

II. Related Work

A. *Information prediction using social network*

The social media sites (e.g., Twitter, Facebook, Weibo) become increasing powerful communication medium for the users to share and interchange the feelings and information of real world events. Becker, Namaan and Gravano [2] explored the Twitter message streams to identify the real and non-real time events using clustering approach over similarity of tweets. It was observed that group communication has a very high correlation with re-tweetability of information [3]. In other research, the social network structure was used to classify and predict the popular domain of discussion or trending topics [4].

B. *Integration of prior work*

Though the prior literature discusses research on information sharing, and event prediction using social networks, there has been a limited research on information-theoretic classification of informed groups by analyzing social, temporal, and second screen device features of group conversations related to TV shows. Ghosh, Surachawala and Lerman [5] classified political campaign based retweeting activity using time interval and user entropy as features. Yang and Counts [6] analyzed information diffusion in Twitter in terms of speed, scale and range using mention ('@') as user interaction but did not include exploration of second screen device features into consideration. Loriche and Coulton [7] investigate the role of social media to facilitate the second screen for TV, thus enabling audience interaction, but they did not examine social or temporal features of interaction. Neither did these prior research considered the measurement of information during group communication via second screens.

III. RESEARCH QUESTIONS

In our research, we identify groups by means of hashtags (#) that can be present anywhere in the comments posted by the TV viewers in Twitter tweets. The collection of viewers with communications messages sharing the same hashtag forms a group. The groups may be small or large depending on the number of interactants within the groups. The probability of information sharing is higher if there are more members within a group. In our study, we express the information shared within a group as a function of social, temporal, and second screen features. The theoretical understanding of our research is characterized by theory of information exchange [11] that delineates an information system as a network of interconnected information nodes, allowing flow of information. The social features deal with the viewers' patterns of conversation within groups and group size. The temporal features identify the volume of real time or non-real time communication. The second screen features include the type of the device people use for group communication. Information sharing in active discussion via second screen by forming groups among viewers lead to our first research question:-

1) Can group interactions using second screen concerning a TV show using second screen be classified in terms of social, temporal, and second screen features?

We term clusters of individuals as informed groups as the members share information about TV show in their group based interactions. The research question is important because commercial and other organizations are increasingly keen to capitalize on prediction of higher informed groups to measure the viewing habits and reactions of the audience.

The first research question eventually leads to exploration of different classification schemes to accurately identify degree of information sharing within the groups. The accuracy of classification of informed groups is different for different classification schemes. In this context, we formulate our second research question as:-

2) Is there a classification scheme that provides best accuracy in predicting the the type pf group interactions using second screen concerning a TV show?

The second research question is important to explore as it involves relative measurement of accuracy to identify the groups that involve high information sharing using second screens. The higher classification accuracy will improve the efficiency of prediction of highly informed groups. It will result in more accurate measurement of viewing habits in terms of sharing of information within groups that can be leveraged by channel owners and other commercial organizations for improved personalization of TV show or the televised commercial products or services.

TABLE 1

ENTROPY BASED CATEGORIZATION OF INFORMED GROUPS

Entropy Range	Class	Group
> 0 to 1.0	0	Least Informed
> 1.0 to 2.0	1	Lower Informed
> 2.0 to 3.0	2	Informed
> 3.0 to 4.0	3	Better Informed
> 4.0	4	Most Informed

IV. DATA COLLECTION AND RESEARCH DESIGN

We select "*Game of Thrones*" as the TV show to evaluate our research questions and collect audience interactions in form of tweets from Twitter. *Game of Thrones* is a popular TV show that is broadcast by channel HBO. The tweet collection is continued with a span from 5th February 2014 to 4th March 2014 and we amassed a total of 317,775 tweets. We use a PHP script to access the Twitter API to collect data concerning the TV show by using the TV show name as the Twitter API query. The tweets retrieved as JSON objects are pulled into a MySQL database. As Twitter is one of the most popular micro-blogging sites, we use it as the platform for the second screen based social interaction. Most micro-blogging services share commonalities such as: a) short text messages, b) instantaneous message delivery, and c) subscription to message updates [8]. So, although we use Twitter as the platform of interaction, we believe that our findings are applicable to other micro-blogging applications.

Once the tweets were collected, we retrieve those tweets filtered by the presence of hashtags (#). In our research, we identify the potential groups by means of these hashtags. In an informed group, there should be at least two or more users interacting with each other and mentioning the same hashtag within tweets. A user can be a member of multiple groups. From our data collection, 4973 informed groups were identified where the range of unique group members is between 2 to 7435.

After formation of the informed groups, we quantify the expected value of the information (i.e., entropy) contained in the tweets posted within a particular group. We use equation 1 to determine the total entropy of a group marked by hashtag *I* using equation 1:

$$H_i = -\sum_j \frac{n_{j,i}}{N} \log\left(\frac{n_{j,i}}{N}\right) \quad (1)$$

Here $n_{j,i}$ is the number of times word j occurs in the collection of tweets containing hashtag i and $N = \sum_j n_{j,i}$ is the total number of words in the collection of tweets containing hashtag i. H_i is the entropy for group marked by hashtag i. The entropy for all 4973 informed groups were computed. We then classified the group conversations into five categories as defined in Table 1. Class 0 is the lowermost informed group, while Class 4 is identified as the highest informed group.

In this study, we extract the social features in terms of count of tweets corresponding to (a) pattern of viewers' conversation within groups, (b) number of unique words present in the tweets, and (c) number of members in a group. The identifiers for categories of patterns for group conversation are described by Table 2.

TABLE 2

CATEGORIES OF GROUP BASED SOCIAL INTERACTIONS

Category	Description
Referral (RF):	Any full length or shortened URL directed at another user. It does not contain any '?' symbol.
Response (RS):	Tweets intentionally engaging another user by means of '@' symbol which does not meet the other requirements of containing queries or referrals.
ReTweet (RT):	Any retweet as recognized by "'RT: @', 'retweeting @', 'retweet @', '(via @)', 'RT (via @)', 'thx @', 'HT @' or 'r @' ".
Broadcast (BC):	Undirected statements (i.e., does not contain any addressing) which allow for opinion, statements and random thoughts to be sent to the author's followers. Any undirected statement followed by questions '?' belongs to Broadcast (BC) category.

We extract temporal features from the perspective of live show timings of the episodes. The episodes of the TV show are broadcasted at 9:00 PM EST every Sunday with duration of one hour, including commercials. There are two such temporal features of the group conversations for a particular informed group: 1) Real-time count: number of interactions recorded when the show is televised live and 2) Non-real time count: number of interactions recorded when the show is not televised live.

Apart from social and temporal features, we explore the second screen device features taking device type into consideration. We classify the second screen device features into three different categories. We leverage the *"Source"* field of the tweets in JavaScript Object Notation (JSON) format. We classify *mobile* tweets based on the existence of keywords within *"Source*" attribute. The keywords we used were "*iPhone", "iPad", "Android", "iOS", "Blackberry"* and *"Mobile Web"*. *"Mobile Web*" refers to access browser based Internet services from handheld mobile devices. If the *"Source"* of a tweet contains any of these keywords, we classify it as a mobile tweet. If the *"Source"* attribute contains the keywords *"TweetDeck"* or *"TweetButton",* we classify the tweet, as *mixed* tweet because the exact platform used to post these tweets is not confirmed. Otherwise, we classify it as a *non-mobile* tweet.

The social, temporal, and second screen device features of an informed group are independent to each other and are used as the predictor variables, while the entropy based categories of the informed groups are considered as response variables. We are using the extracted feature set to predict the class of an informed group to identify whether it corroborates more information.

TABLE 3

CATEGORIES OF GROUP BASED SOCIAL INTERACTIONS

	Class				
	0	**1**	**2**	**3**	**4**
Training	42	472	1628	874	307
Test	21	238	801	438	152

V. Methodology

To investigate our research questions, we evaluate different classification algorithms for class prediction. The size of data (i.e. the set of response and predictors) is varied across the classes. Prior to applying the classification schemes, use principal component analysis (PCA) to identify the contributing features in determining the entropy based class of informed groups. Once the feature set is selected, the data for each class is segregated into 2:1 ratios, thus form a training set with 67% of total data and test set with the remaining 33%. The training set is formed by randomly chosen response-predictor variables from each class. The distributions of training and test set across five classes are given in Table 3.

The family of classification schemes that is selected in this research includes six algorithms listed as: 1) Multiclass Linear Discriminant Analysis (mLDA), 2) Quadratic Discriminant Analysis (QDA), 3) Multinomial logistic regression, 4) K-nn classification, 5) Decision tree classifier and 6) multiclass Support Vector Machine (mSVM).

TABLE 4

CATEGORIES OF GROUP BASED SOCIAL INTERACTIONS

Input feature	Contribution
Response (RS)	1.113
Referral (RF)	1.964
Retweet (RT)	1.353
Broadcast (BC)	1.647
User count (UC)	1.572
Real-Time Count (rTC)	0.019
Non-Real Time Count (nrTC)	1.520
Unique Words (UW)	1.997
Mobile device (MD)	1.414
Non-Mobile device (nMD)	1.656
Mixed device (MxD)	1.191

VI. Result

Before applying the classification schemes, we need to find out and retain the predominant features that contribute most in defining the response variable. We perform PCA to find out those prominent features. We use covariance matrix for eigenvalue decomposition of data and find out that first two components explain 99% variance of the data as shown in Figure 1. We adopt the algorithm described in previous research [9] to identify the prevalent features by calculating the contribution of the features to the eigenvalues of these two components. Table 4 displays the sum of entries (i.e., contribution) of each input feature over these two eigenvectors. From the contributions observed in Table 4, a subset of ten features is selected. These features are used as predictors for the classification algorithms. The set excludes Real-Time Count (rTC).

A. *Multiclass linear discriminant analysis (mLDA)*

We carry out multi-class linear discriminant analysis over the training data with equal prior probability of 0.2 for each class (See Table 1). In our experiment mLDA gives an overall misclassification rate of 0.275 of the test set.

B. *Quadratic discriminant analysis (QDA)*

We apply QDA over our training set with the equal prior probability (0.2) for each class. In QDA, the multi class classification is done by means of a quadratic surface. In our research, the overall misclassification rate for QDA is 0.468.

C. *Multinomial logistic regression*

As we have more than two classes related to the degree of information sharing among the informed groups, we use the entropy based class membership of the training data as categorical response in multinomial logistic regression given the input features as independent predictors. The experiment with multinomial logistic regression yields overall misclassification error of 0.210.

D. *K-nearest neighbor*

Our training examples are vectors in a multidimensional feature space each with a class label. We apply k-nearest neighbor algorithm on the training set as one of the classification schemes where an item is placed in a class most common among its k nearest neighbors by majority vote of its neighbors. The k value with minimum overall misclassification for prediction of test set is selected. Figure 2 displays the misclassification errors for the test set over different user specific k values range from 1 to 11. The metric for distance function between two neighbors is considered Euclidean distance. From Figure 2, it is observed that k-nn algorithm gives minimum misclassification rate of 0.039 for k = 5. So, 5-nn classifier is the best choice in our research.

E. *Decision tree classifier*

Decision tree is a non-parametric supervised learning algorithm with a goal to predict the value of a target variable by learning simple decision rules inferred from the data features using recursive partitioning of source space. We use decision tree classifier with a multiclass framework supported in R. The model is trained over the training set and gives a prediction misclassification rate of 0.084.

F. *Multiclass support vector machine (mSVM)*

SVMs are inherently two-class classifiers. Crammer and Singer [10] extend it with multiclass formulation by building multiple binary class classifiers with either i) one-versus-all (OVA) or ii) one-versus-one (OVO) strategy. We use mSVM to build five classifiers using one versus all (OVA) strategy where a single classifier is trained per class to distinguish that class from all other classes. We use radial kernel with cost as 1000 and gamma as 1. It was found that overall prediction misclassification for the test set is 0.176.

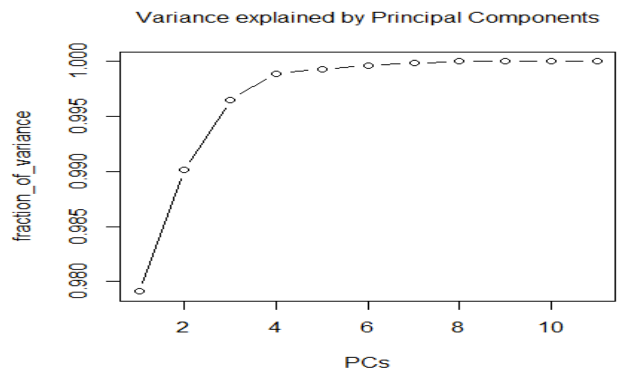

Fig. 1. Cumulative variance explained by principal components

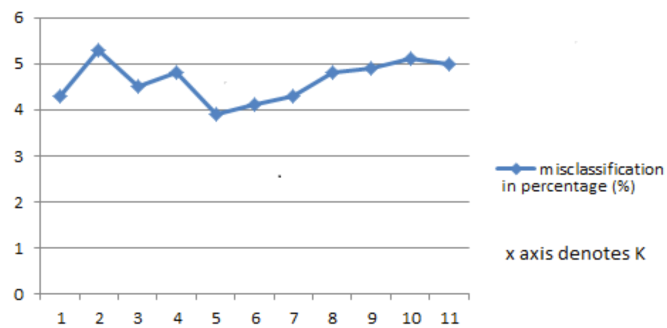

Fig. 2. K values vs. misclassification error (%)

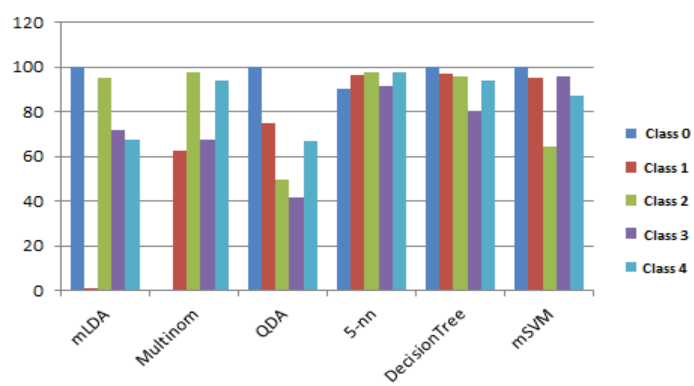

Fig. 3. Intra-class correctness of prediction (%) for all classifiers

We further look into the distribution of misclassification across five response categories for all classification schemes used in our study. From the five different confusion matrices we constructed Figure 3 that illustrates the relative correctness for prediction over test data done by the six classification algorithms across five categorical response variables. 5-nn classifier predicts best for most of the classes with an average classification among all classes of more than 90%, while mLDA and QDA though work best for class 0 data; their prediction accuracies for high-informed groups are very poor.

The analysis satisfies the first research question by identifying the relationship between entropy-based categories of informed groups and the input feature set. The second research question is also addressed as the 5-nn classifier gives the best prediction accuracy from the perspective of minimum overall misclassification rate.

VII. Discussion and Implication

While investigating the first two research questions, we find that there is a relationship between the entropy based classification of the informed groups and the social, temporal, and second screen device features obtained from the collected tweets. The existence of relationship and the identification of best classification approach address two research questions. We further explore the intra-class prediction across all six classification schemes to find the distribution of prediction efficiency over low and high informed groups. It is observed from Figure 3 that 5-nn classifier shows greater than 90% prediction correctness across all classes of informed groups followed by Decision tree with more than 90% accuracy for all classes except class 3. mLDA ,and QDA classifiers work perfect for the lowest informed group but perform inefficiently in determining the higher informed groups. mSVM classifier shows very good accuracies for two lowermost and two uppermost classes but doesn't work well for class 2. Multinomial regression fails for the lowest informed groups.

A. *Theoretical Implication*

Concerning the theoretical implications of the findings, the understanding of theory of information exchange [11] is characterized by the proposed entropy based classification of the group of viewers that is treated as information systems with viewers as information nodes. The information or feeling is shared and interchanged within the system using social media platforms as communication channels.

B. *Practical Implication*

Accurate identification of informed groups where members interchange views about TV show content and televised ad will inevitably assist content providers and retailers to extend their business opportunities. We have classified the informed groups into five categories that range from lowest informed group to highest informed group based on entropy of information exchange within a group. The discussion about the TV shows and advertised products or services in high informed groups always have greater probability of information diffusion than that in lower informed groups. From a retailers' point of view, the information about the brand might be diffused to the viewer's larger social community, while channel owners may have greater understanding about group member's feeling and reaction about the show content. So, broadcasters and marketers should tap in these higher informed groups to explore the stronger message association from the perspective of personalization of TV show and higher purchase intent among the potential customers of the brands. Both broadcasters and retailers can monitor the informed groups mainly in non-real time as contribution of real Time Count feature (rTC) is negligible compared to non-real Time Count (nrTC) (See Table 4).

VIII. Conclusion

In our research, we analyze the relationship between entropy-based categories and the input features (social, temporal and second screen decice features) of a TV show related second screen communication done in groups that are formed by hashtags. We observe that 5-nn classification provides the best prediction accuracy in identifying the informed groups of viewers. The accuracy of identification of different informed groups enable the retailers and channel owners to create new business strategies by tapping the ad and TV show related information sharing wihin different types of groups.

In our research, we don't evaluate the prevalent second screen communication pattern in group conversation. Neither have we explored the primary second screen for group interaction in terms of device usage. We will evaluate the effect of secondary screen in terms of linguistic pattern and usage of device during group based interaction exchange in future research. In present study only one TV show is considered for research. We will extend our research over several TV shows for the generalizability of our findings.

References

[1] P. Mukherjee and B. J. Jansen. (2014). Social TV and the social soundtrack: significance of second screen interaction during television viewing. Presented at SBP, Springer, pp. 317-324.

[2] H. Becker, M. Naaman and L. Gravano. (2011), Beyond Trending Topics: Real-World Event Identification on Twitter. Presented at ICWSM, pp. 438-441, AAAI Press.

[3] B. Suh, L. Hong, P. Pirolli and E. H. Chi. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. Presented at IEEE trans. on Social Computing, pp. 177-184.

[4] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A.Choudhary. (2011). Twitter trending topic classification. Presented at IEEE trans. Data Mining Workshops (ICDMW), pp. 251-258.

[5] R. Ghosh, T. Surachawala and K. Lerman. (2011). "Entropy-based Classification of Retweeting Activity on Twitter", arXiv preprint arXiv:1106.0346, unpublished.

[6] J. Yang and S. Counts. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter", Presented at ICWSM, pp. 355-358, AAAI Press.

[7] M. Lochrie, and P. Coulton. (2011). Mobile phones as second screen for TV, enabling inter-audience interaction, Presented at International Conference on Advances in Computer Entertainment Technology (ACE), pp. 1-2.

[8] B. J. Jansen, M. Zhang, K. Sobel and AChowdury. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology, 60(11)*, pp. 2169-2188.

[9] F. Song, Z. Guo, and D. Mei. (2010). Feature selection using principal component analysis, System Science, Engineering Design and Manufacturing Informatization. Presented at ICSEM, pp. 27-30.

[10] K. Crammer and Y. Singer. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research, 2*, pp. 265-292.

[11] S. K. Chang. (1982), Information exchange theory and man-machine interaction, Presented at IEEE Trans. on Decision and Control, 21, pp. 583-588.