2015

# Extracting Greek Elections Tweet's Characteristics

Katerina Kalaidopoulou
*Athens University of Economics and Business, Greece*, kkalaid@gmail.com

Angeliki Romanou
*Athens University of Economics and Business, Greece*, agromanou@gmail.com

George Lekakos
*Athens University of Economics and Business, Greece*, glekakos@aueb.gr

# EXTRACTING GREEK ELECTIONS TWEET'S CHARACTERISTICS

*Completed Research*

Kalaidopoulou, Katerina[1], Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece, kkalaid@gmail.com

Romanou, Angeliki[1], Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece, agromanou@gmail.com

Lekakos George[1], Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece, glekakos@aueb.gr

## Abstract

*Social media offer platforms that anyone can use, giving the opportunity to share information among networks in an easy and interactive way. It is not a surprise that social media marketing has become a primary focus on both digital and traditional revenue models of businesses. In this work, information sharing by users in the context of Twitter is studied, by modeling message's characteristics and users' behavior about Greek 2015 January elections. A detailed data set about tweets' characteristics such as length, existence of URLs or hashtags and mentioning of other users, is collected after the elections day, and the relationships between related users and network's responses on the shared tweets, are examined. An unsupervised clustering model is implemented on tweets' characteristics using CRISP-DM methodology. The empirical results suggest the existence of different content groups, such as tweets with extensive text, URLs and hashtags which can be characterized as "Linked" type of shared content.*

Keywords: Machine learning, social media analysis, tweet clustering, Twitter

## 1. Introduction

Social media have become the most popular way to create, share and exchange information, pictures, music, videos, thoughts in digital communities. Facebook, Twitter, YouTube, LinkedIn are just a few of them. As people follow, like, share, tweet, retweet, tag, rate and text one another, they become part of an enormous social network, providing the opportunity to be extracted and analyzed towards identifying users' behavioral patterns and performing more effective information sharing and diffusion. The identification of the critical characteristics of the messages, that enables the maximization of its impact in a network, is a critical business goal and important users' need which paves the way for extensive research from both industry and academic community.

This need is increasing when it comes to analyze collected information under government elections' context. Related studies on this matter have been conducted on many different

[1] ELTRUN: The e-Business Research Center

topics such as elections, crises, televised events (Bruns and Stieglitz 2012). According to social sciences' studies, user has a great influence on the message's impact on the network. However, while the message, i.e. tweet, itself may influence the response it will get, tweet's characteristics, such as length tend to be neglected when mining them. In this work we focused on using a content-based approach on grouping messages/tweets on a time frame of a week after Greek January elections which are considered to be one of the most critical and diverged elections for Greece in the last fifty years. Extracting groups of similar tweets based on their characteristics (length, existence of URLs etc.) can lead to better understanding of the different types of tweets, correlating each of their characteristics with the network's response.

We contribute to this field by exploring Twitter social network so as to identify different types of tweets based on their characteristics such as length, hashtags etc. Which types of posted tweets maximize the network's response? How can we predict which characteristics are most related to high response and alternatively indicate high impact on the network? This research aims to approach these questions and pave the way for extensive future work on content characteristic-based analysis. We also seek to introduce an alternative way of tweets' segmentation, based on their characteristics and correlate them with users' attributes and network's response.

# 2. Research background

Previous studies have explored how social media users post, reply or forward messages, i.e. in Twitter how they tweet, retweet and reply. A more specific study conducted by Boyd et al. (2010) analyzed how, why and what people retweet in Twitter network, concluding that Twitter is mainly seen as a conversational environment. Though, in the context of user-centric classification, previous research has indicated that weak ties (in the form of unidirectional links) are more likely to engage in the social exchange process of content sharing. (Shi et al, 2013). Moreover, Pennacchiotti, Marco, and Ana-Maria Popescu (2011) have attempted to classify users based on a comprehensive set of features derived from such user information. Additionally to this research, Stieglitz, Stefan, and Linh Dang-Xuan (2013) seek to explain "whether sentiment occurring in social media content is associated with a user's information sharing behavior" (2013).

Many scholars have also studied Twitter activity in the context of individual case studies, which represent many different topics, such as elections, crises, televised events etc. A more generic study, conducted by Bruns and Stieglitz (2012) implies that "standard response to the emergence of breaking news and other acute events is the tendency to find, share, and re-share relevant information, resulting in a high rate of URLs and retweets". On the other hand, in live and streaming events "Twitter acts as a backchannel, containing mainly original commentary that does not engage with the tweets of others or provide a substantial number of links to further information".

As a result, it is indicated that in different types of tweets, different characteristics may drive awareness and response from network's peers. Focusing more on elections, we found that other researches cover respective events from other countries in terms of predicting the elections' results (Tumasjan et al, 2010) or analyzing public communication (Bruns and Burgess, 2011). In our case we try to identify the most important characteristics of the message that maximize the response it will get from the network. Literature has issued the need for further content-based analysis on social media published material. However, none of the so-far work has approached content analysis on the context of Greek 2015 controversial elections. Thus, we contribute to this field by performing an exploratory analysis based on tweets' content characteristics, deriving distinct clusters of tweets based on tweets' and their publishers' attributes. We are using the CRISP-DM methodology (Chapman et al, 2000), a concrete methodology for unsupervised machine learning, whose phases are analyzed below.

## 3. Methodology and Data

Our analysis process was implemented according to the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for approaching data mining problems. In order for the methodology to meet this work's goals, we implemented four out of five steps of Crisp DM process (Figure 1) - Data Understanding, Data preparation, Modeling and Evaluation - which will be extensively analyzed on the following section.

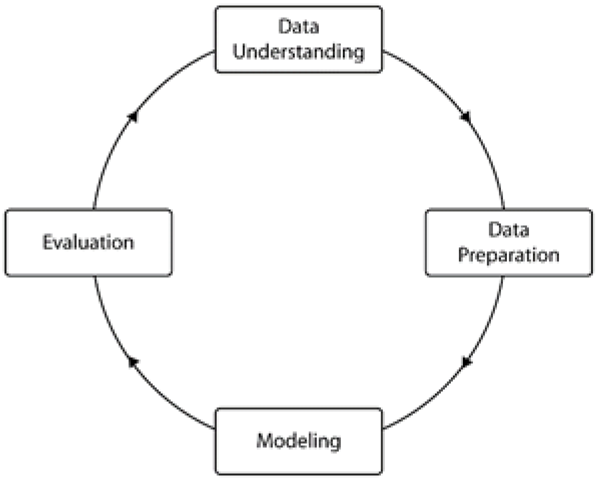


*Figure 1: CRISP-DM methodology process*

In order to initiate the aforementioned data analysis process, we collected network data using the NodeXL data importer, which allows access to social media and other forms of networks. In our analysis we imported data from "Twitter Search Network" because it allows the extraction of networks according to one or more hashtag(s). We used NodeXL client software to download Twitter data and selected only tweets that contain the hashtag "ekloges2015" which means elections of 2015 in Greek language, on a time frame of one week after Greek elections. The collected dataset consisted of the basic network, which includes published tweets and users' characteristics - such as user's "follows" or "mentions" or "replies" - along with the user's network of friends and the relationships between them. In order to achieve this research analysis' goal, additional data variables had to be used. Giving the limitations of NodeXL to get access on tweets' Retweets, Favorites and Comments, we directly requested data from Twitter's API. Using the Twitter's API console, we downloaded for the same date range and the same hashtag the additional variables: Retweets, Favorites and Comments (Replies) for each observation.

# 4. Analysis and Results

## *4.1. Data Understanding*

The data understanding phase starts with an initial data collection and proceeds with activities which aim to become familiar with the data, identify data quality problems, discover first insights and detect interesting subsets in the dataset. In the extracted network, vertices represent user entities while edges represent the interaction between them (Figure 2). Such interaction can be either a "follow" relationship between two users, a user "mention" in tweets, or a user "reply" to a tweet. Posted tweets are represented in the graph by a self-loop on the user who posted it. A "follows" edge means that one user follows another in the selected network. A "mentions" edge is created when one user mentions another user in a tweet (e.g. "being in the conference with @someone"). A "reply" edge is subtype of "follow" because it labels a relationship when one user refers to another at the beginning of the tweet (e.g. "@someone speaking right now"). Finally, a tweet is a simple post without a "reply" or "mention". The entire network consists of 852 vertices and a total of 28.866 edges between them, in which 242 are tweets, 27.694 regard "follow" relationships and 930 are "mentions" to the tweets.

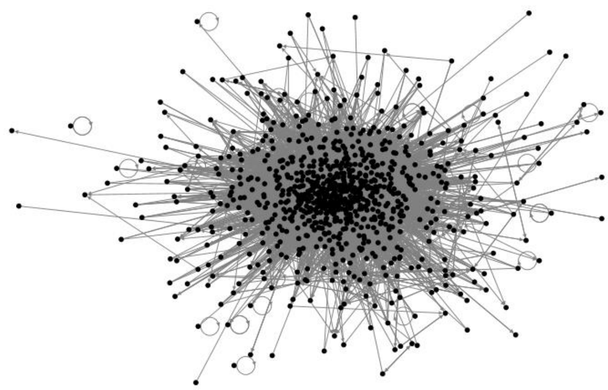

*Figure 2: Extracted Twitter network*

The structure of the initial data set we exploited, combines data from the two data sets and follows an entity - based approach. Thus, there are three distinct entities: "Users" who post content, "Tweets" which are the posted messages on Twitter and "Response" which is described by the actions users took on the posted tweets. Merging graph data with Twitter's API export results, we composed a dataset with three different kinds of variables related to those entities. These data need to be exploited in order to create the useful for the analysis meta-data, as described in the following section. Our goal is to extract variables describing tweets characteristics - such as length, number of additional hashtags, number of URLs etc. -, attributes for user characteristics - such as number of followers, date of registration, total number of posts etc. - and data related to the response that each tweet got on Twitter - such as Retweets, Favorites and Mentions.

### *4.2. Data Preparation*

The data preparation phase covers all activities to construct the final dataset from the initial raw data. As mentioned before, NodeXL provides the tweets - comments that are neither replies nor mentions- and some basic descriptive statistics such as the exact hashtags and URLs used. However, our research goal is to extract more useful meta-data like the tweets' length, number of URLs attached, number of hashtags used and number of mentions of other users. Thus, and within the entity - based approach, we isolated the tweets from NodeXL and the respective user's name that posted them.

Then we created the "Tweets Table" to store them and their most important attributes. Using appropriate SQL queries, we derived the following attributes: Tweet id (numeric), Content length (numeric), Number of hashtags of the tweet (including #ekloges2015), number of additional hashtags (other than #ekloges2015), number of attached URLs, number of mentions of other users. Merging the tweets' data with the Twitter's API export results, we expanded the dataset with three different kinds of variables related to those entities. Next, for each user that posted a tweet in this table, we isolated the useful data about them. These attributes are ID, Number of people that they Follow, Number of Followers, Number of Tweets posted, Number of Favorites.

According to the analytical approach that we chose to implement in this work, the dataset that will serve as input on the decided model should be properly be transformed. Our approach is to perform clustering analysis on tweets characteristics. To do so, we needed data for each tweet on whether it has a specific attribute or not. We coded this information into binary variables in where the value "1" represents the presence of this attribute in a tweet, and the value "0" represents its absence. However, we noticed that the tweets' variables were neither at the same scale nor on the same type in order to transform them into binary variables. To overcome this issue, we manipulated those attributes by scaling the numerical variables into classes. Then we either assigned the tweet on each class (value: 1) or not (value: 0). For example, the numerical attribute "Length" had a range from 35 characters (min) to 145 characters (max). We scaled this variable into 6 classes. Thus, in the final dataset a tweet that had 60 characters length was stored in the dataset in the following format:

| TweetID | length_class [35-53] | length_class [54-71] | length_class [72-89] | length_class [90-107] | length_class [108-125] | length_class[1 26-145] |
|---|---|---|---|---|---|---|
| 100 | 0 | 1 | 0 | 0 | 0 | 0 |

### *4.3.Modeling*

Having prepared the data and defined the fact table, we proceeded with the implementation of the clustering model. Our goal is to identify clusters of tweets' characteristics, meaning tweets that have common characteristics and thus can be grouped together. From our final dataset we used only the binary columns that represent tweets' attributes and consists of all types of data in the transformed classes which serve as the input of the model.

However, our research is based on a binary fact table so as to enhance the understanding of the results. Using binary data enables the researcher get results and understand each attribute value separately from the other values. In order to build the model, we used the x-means algorithm, which uses heuristically the k-means algorithm and can define the number of the deriving clusters without user's input. The algorithm uses a heuristic approach when creating the model and thus it is not a prerequisite to give the number of clusters as input. More specifically, the algorithm runs the model with different numbers of clusters on each iteration and ultimately selects the optimal number of clusters that maximizes the model's information gaining. At the end, it returns the possibilities of each attribute or attribute's value to appear in the respective cluster.

Implementing clustering algorithm on this dataset, we derived three clusters. The biggest cluster was "Cluster 1" accounting for the 68.95% of the total dataset, followed by "Cluster 2" which accounts for the 30.24% of tweets and "Cluster 3" which covers the remaining 0.81%. The strongest relation is shown between "Cluster 1" and "Cluster 3" in comparison with "Cluster 2". On the following section, we continue with the evaluation phase of the CRISP-DM methodology identifying more descriptive statistics about the clusters and concluding to some clusters' characterizations.

### *4.4. Evaluation*

As the first step of the evaluation phase, we identified the main characteristics of each cluster. The clustering results provide us with the possibility of each attribute to appear in the tweets of the respective cluster.

***Cluster 1:*** The main characteristic of the tweets in "Cluster 1" is that they all have at least one URL attached (100%). They also have 1 or 2 hashtags more that the examined one (#ekloges2015) (48%). Finally, the length of the tweet is from 126 up to 145 characters (46%), which means that they belong to "Big-sized" tweets. However, most of the characters in such tweets belong to the attached URL and not the message itself. In addition, only 1% of the tweets may "mention" another user. As a result, we label tweets of "Cluster 1" as "**Linked**".

***Cluster 2***: The first half of the tweets in this cluster has only the #ekloges2015 hashtag (51%), while the other half (47%) has 1 or 2 more hashtags. Furthermore, the tweets are medium (22% within the range 108 - 125 characters) to big sized (38% has from 126 up until 145 characters), which means that they have about 120 characters on average. Having no URLs attached and without mentioning other users, these tweets are characterized as "**Linguistic**".

***Cluster 3***: Tweets of this cluster have no other hashtags than the #ekloges2015 one (100%). In addition, more than half of them have a URL attached to the tweet (58%) and they are medium sized in length containing 72 to 89 characters (58%). Moreover, there is no mentioning of other users and thus these tweets are called "**Focused**". However, this cluster contains only few tweets and thus it is statistically insignificant to continue with further analysis in it.

These characteristics are shown in the following figure (Figure 3):

| Variables | Population (All) | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| | | Multimedia | Linguistic | Focused |
| Length (108-125) | 42 | 32 % | 22 % | 0 % |
| Length(126-145) | 65 | 46 % | 38 % | 0 % |
| Length (35-53) | 5 | 1 % | 0 % | 21 % |
| Length(54-71) | 15 | 0 % | 15 % | 21 % |
| Length (72-89) | 19 | 5 % | 6 % | 58 % |
| Length (90-107) | 26 | 17 % | 18 % | 0 % |
| #Hashtags(0) | 91 | 42 % | 51 % | 100 % |
| #Hashtags(1-2) | 73 | 48 % | 47 % | 0 % |
| #Hashtags(3++) | 10 | 10 % | 3 % | 0 % |
| Mentions | 1 | 1 % | 0 % | 0 % |
| URLs in Tweet | 92 | 100 % | 0 % | 58 % |

*Figure 3: Clusters' characteristics*

After tweets clusters' characterization, we seek to understand what kind of response the tweets of each cluster received. As no tweet within the examined period got a reply, we will assess only the retweets and the favorites. In total, tweets in "Cluster 1" have collected 185 retweets and 149 "favorite" labels. On the other hand, tweets in "Cluster 2" have 97 retweets and 78 "favorite" labels respectively.

However, the clusters do not have the same size and hence cannot be objectively compared. As a result, we calculated the average number of retweets and favorites a tweet may get in each cluster. According to the results, a tweet that belongs to "Cluster 1" receives on average 1.16 retweets and 0.93 "favorite" labels, while a tweet in "Cluster 2" may get 1.33 retweets and 1.10 "favorite" labels. Interpreting the results, we can assume that "Linguistic" tweets are more likely to get a better response from their network, than "Linked" tweets.

The next step of our analysis focuses on the types of user profiles that post the tweets in each cluster. The user attributes that are available in our data set and represent the user's activity and behavior in Twitter are the number of people the user followed, the number of users that follow the user, the number of tweets the user posted in its entire history and the number of favorites it has. From these four attributes, we focus mainly on the total number of tweets the user tweeted, because it represents the intensity of the user's activity and the number of followers the user has, because it represents the maximum range of influence.

Users from "Cluster 1", i.e. users who posted tweets that belong to "Cluster 1", have a big number of followers and activity (i.e. tweets). However, users in "Cluster 2" exceed the first ones, by being very active and having the greatest number of posted tweets and followers among all three clusters. Moreover, we calculated the average of each metric and concluded with the following results. Users who post "Linguistic" tweets are the most active and followed, as they have the greatest number of tweets and followers, while users who post "Linked" tweets are the second most followed.

## 5. Discussion

This research is an exploratory work which aims to identify tweets' attributes related to its content, group them together according to their common characteristics and drill information about the impact of the tweet on the network. Using the tweets' characteristics, the tweet's impact on the network and the user's profile attributes, we aim to relate the clusters with the possible response they might get from the network. In addition, we aim to relate them with the types of users that post tweets of each cluster, in order to reach to some conclusions about the tweets' characteristics that may lead to bigger response rate for a typical Twitter user.

The results show that tweets can be grouped in two distinct clusters with unique characteristics that differentiate them from other groups. Some tweets focus on promoting a link to a video, a picture or another website ("Linked") and adding more hashtags to become visible, while others focus on communicating a message through a medium-length text in a tweet ("Linguistic") and mostly using one hashtag. Of course, there are also those tweets that point out their message via a small but targeted tweet ("Focused"), but they have not been examined further. Each type of tweet serves a different purpose, but also has a different impact on the network.

From the results, we derive that Twitter users tend to retweet more "Linked" tweets because they may contain interesting information and want to share it with their network. Such tweets reach many and sometimes different topics through the usage of more than one hashtags. On the other hand, they tend to favorite tweets with a long message, because they may be more meaningful for them. These results imply that if someone wants to be heard on the network, they have more possibilities to succeed if they accompany the message with a link to an interesting content, so as to indulge other users to share it with their network through retweets. This may be an effective way to make the message viral on the network. However, if someone wants to engage with their network, they should post long but meaningful content in order to entice others to favorite them. The results do not favor targeted but short tweets, which are neither favored nor retweeted. However, this cluster contains only a few tweets and thus no concrete conclusions can be derived.

The above-mentioned findings are enriched with some insights about the users who post different types of tweets. The most active users in terms of number of tweets post "Linguistic" tweets, while the most favored ones post "Focused" tweets and the most followed and the second most favored ones post "Linked" tweets. These results shed light from a different perspective as they clearly imply that the most followed users do not post "Linked" tweets, but tend to post medium-sized tweets in terms of content length and be favored more times. This could mean that they prefer to stay focused on one topic by using only one hashtag, promote one concrete message without the distraction of a link and build engagement with the network. We supposed this behavior is explained by the fact that known users with loyal followers do not need to use URLs, interesting articles and many hashtags to catch network's attention and create buzz.

Examining further these results, we looked in our data set to see, what type of tweets did the most famous users tweet. By famous or known users we refer to twitter accounts that represent influential individuals – like politicians, political analysts or journalists who are popular on society – and commercial accounts -like newspapers or blogs that became popular through digital social media. In order to achieve this, we identified the users that represent newspapers, famous blogs and news agencies in general. Such users tend to post more tweets that belong to Cluster 2, i.e. medium sized posts, with few hashtags and no URLs. However, there are also many online news agencies that post tweets mainly with URLs attempting to advertise their blog or website, and thus belong to Cluster 1. As a result, we concluded that there are no clear trends from the famous users.

# 6. Limitations - Future research

NodeXL Twitter Search network operator communicates with Twitter service in order to import data. Our query targeted 1000 tweets which were easy to be found because of this hashtag's popularity (our hashtag was #ekloges2015 and the data collection happened right after the Election Day). However, we imported not only the tweets, but also the network for each user who posted the tweet. Thus, NodeXL downloaded only 250 tweets, a relatively small dataset, setting the first limitation of our research. A small dataset narrows the scope of our research and can be biased as it is not representative of the total population. Future research will examine a different dataset, with larger number of tweets which will be extracted using the Twitter API. Using the latter method we could target specific fields (e.g. tweet, tweet_id, user, number of followers etc.) which are needed for our research and belong to a predetermined time frame. Extracting less but more useful information would lead to collecting a much bigger dataset in less time.

Our research focuses on tweets about a specific occasion, the Greek national elections of 2015. Hence the results are applied only to this topic and we cannot assure external validity to any other type of event, occasion or topic. In addition, tweets were extracted from a timeframe of seven days after the elections had finished and thus do not represent users' behavior throughout the elections campaigns when tweets may get different response - more retweets or replies. Additional work could compare analysis results from both before elections date and after. More specifically, our next step is to select a specific elections' campaign beforehand, so as to collect tweets a few weeks before the elections, throughout the elections' day and a few weeks afterwards. In this way, we will cover the whole elections' period and be able to make comparisons and more in depth analyses.

Moreover, our current study uses exploratory method on social network analysis through a heuristic approach, using k-means, a heuristic algorithm, which is not without its limitations. First of all, it cannot work with categorical data, but only numerical values (Huang, Zhexue, 1998). In addition, another limitation of this clustering algorithm is that it cannot handle empty clusters and outliers, while the researcher has to reduce SSE with post-processing tasks (Singh et al, 2011). Future studies could use another clustering algorithm to compare the results or switch to a different analysis approach. We are currently extending this work by switching on explanatory analysis aiming to identify the most important tweets' attributes that lead to intense response from the network and how we can predict response outcome based on tweets' and users' characteristics.

Another limitation derives from the scope of the current research, which focuses on the contribution of tweet's attributes on the impact a tweet has on its network. However, social sciences show that user can affect tweet's impact. Thus, in future studies we will examine if the popularity of the user may influence the response a tweet has on its network. In order to define objectively the term 'popularity', we will extract tweets from the top tweeters who used the relevant hashtag. Comparing their impact to the less 'popular' users' impact, we could enhance our analysis and answer our hypothesis.

Moreover, this study examines the impact of shared content (tweet) in Twitter, which is only one from many ways to interact with a network in available social media platforms. Future research could compare which attributes relate to the influence of a message in different social platforms. In specific, we could extract posts (e.g. tweets from Twitter, posts from Facebook etc.) about the same topic and compare their attributes among different platforms. Finally, in addition to the above-mentioned findings, it would be interesting to identify whether the big network of followers leads to more retweets of the message or the many retweets lead to a bigger network of followers.

## 7. References


Boyd, danah, Scott Golder, and Gilad Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." HICSS - 43. IEEE: Kauai, HI, January 6

Bruns, Axel, and Stefan Stieglitz. "Quantitative approaches to comparing communication patterns on Twitter." Journal of Technology in Human Services 30.3-4 (2012): 160-185.

Bruns, Axel, and Jean Burgess. "# ausvotes: How Twitter covered the 2010 Australian federal election." (2011): 37.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Shearer, C. "1.0 Step-by-step data mining guide, CRISP-DM Consortium, 2000."

Cheung, Kwok-Wai, et al. "Mining customer product ratings for personalized marketing." Decision Support Systems 35.2 (2003): 231-243.

Giudici, Paolo, and Gianluca Passerone. "Data mining of association structures to model consumer behaviour." Computational Statistics & Data Analysis 38.4 (2002): 533-541.

Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.

Griva, Anastasia, Bardaki, Cleopatra, Panagiotis, Sarantopoulos, Papakiriakopoulos, Dimitris,, "A data mining-based framework to identify shopping missions" (2014). MCIS 2014 Proceedings.

Harrison, Tina S. "Mapping customer segments for personal financial services. "International Journal of Bank Marketing 12.8 (1994): 17-25.

Huang, Zhexue. "Extensions to the k-means algorithm for clustering large data sets with categorical values." Data mining and knowledge discovery 2.3 (1998): 283-304.

Kwan, Irene SY, Joseph Fong, and H. K. Wong. "An e-customer behavior model with online analytical mining for internet marketing planning." Decision Support Systems 41.1 (2005): 189-204.

Pennacchiotti, Marco, and Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." ICWSM 11 (2011): 281-288.

Shi, Zhan, Huaxia Rui, and Andrew B. Whinston. "Content sharing in a social broadcasting environment: evidence from twitter." Available at SSRN 2341243 (2013).

Singh, Kehar, Dimple Malik, and Naveen Sharma. "Evolving limitations in K-means algorithm in data mining and their removal." International Journal of Computational Engineering & Management 12 (2011): 105-109.

Stieglitz, Stefan, and Linh Dang-Xuan. "Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior." Journal of Management Information Systems 29.4 (2013): 217-248.

Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." ICWSM 10 (2010): 178-185.