8-2018

Detecting Popularity of Ideas and Individuals in Online Community

Chien-Yi Hsiang
Purdue University

DETECTING POPULARITY OF IDEAS AND INDIVIDUALS IN ONLINE COMMUNITY

by

Chien-Yi Hsiang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Purdue Polytechnic Institute

West Lafayette, Indiana

August 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Julia M. Rayz, Chair

Department of Computer and Information Technology

Dr. Kathryne A. Newton

Department of Technology Leadership and Innovation

Dr. John A. Springer

Department of Computer and Information Technology

Dr. Victor Raskin

Department of English and Linguistics

Approved by:

Dr. Kathryne A. Newton

Head of the Graduate Program

To my Heavenly Father

ACKNOWLEDGMENTS

I wish to express my most profound gratitude to my advisor, Dr. Julia M. Rayz. This dissertation would not have been possible without her mentorship. Julia always puts the students' need in the priority and does her best to help us unconditionally. Julia is always available when I have questions, but patient and understanding when I have difficulties, giving me enough time to learn and grow. Julia is someone you know you can always turn to seek advice and rely on. I feel blessed to be one of the students of such a wonderful advisor.

I am also very grateful to all the members in my Dissertation Committee, Dr. Victor Raskin, Dr. Kathryne A. Newton, and Dr. John A. Springer, for their extensive feedback and valuable guidance. I am especially indebted to Dr. Newton because she has been supportive of my academic decisions and helped me with the assistantship to continue pursuing my Ph.D. degree.

I would like to thank all my colleagues in the Applied Knowledge Representation and Natural Language Understanding Lab at Purdue University. I appreciate all their helpful input and comments on this dissertation. I would especially like to thank Qiaofei Ye and Shih-Feng Yang for their technical supports; Kanishka Misra for his aid in the data visualization; as well as Gilchan Park, Shannon Jing, and Parag Guruji for the brainstorming time. I had the pleasure to learn from all of you.

I have had the great fortune of loving support and help from my friends on this journey. Heartfelt thanks to Yunxia Pang, Zhenyu Tang, Chi-Shia Hsu, Tintin Wei, Kuanling Lin, Lisa Tam, Emma Yang, Lan Jin, Yingying Hong, Sooyeon Choi, Eonyoung Cho, Charles Song, Roy Luan, Edith Lin, and Hanxi Sun. All of you have helped me in some way during my pursuit of Ph.D.

Finally, I would like to thank my parents, whose infinite love and support have been always with me. My grandfather passed away few months before my final defence, but I know he must be very proud of me if he were here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

- **Idea Adoption:** Some of the submitted ideas in *MyStarbucksIdea* online community were adopted by Starbucks to be Starbucks products or services.
- **Idea Embedding:** Representing the overall meaning of the submitted ideas in the *MyStarbucksIdea* online community through Doc2Vec
- **The Success of Ideas:** The ideas adopted by Starbucks
- **The Popularity of Ideas:** High comment number from other users
- **The Popularity of Individuals:** High indegree centrality
- **Users' Online Behaviors:** Users' online activities in the *MyStarbucksIdea* online community, including idea posting, commenting on other's ideas, commenting on their own ideas, voting (like or dislike) about ideas.
- **SMOTe + Tomek Links:** The combination of the Synthetic Minority Over-sampling Technique (SMOTe) with Tomek links.
- **SMOTe + ENN:** The combination of Synthetic Minority Over-sampling Technique (SMOTe) with Edited Nearest Neighbor (ENN).

LIST OF ABBREVIATIONS

- AUC: The area under ROC curve
- CNN: Convolutional Neural Networks
- Doc2Vec: Document to vector, also known as paragraph vector
- OLS Regression: Ordinary Least Square Regression
- RNN: Recurrent Neural Network
- SD: Standard Deviation
- SVM: Support Vector Machine
- NLP: Natural Language Processing

ABSTRACT

Author: Hsiang, Chien-Yi. PhD
Institution: Purdue University
Degree Received: August 2018
Title: Detecting Popularity of Ideas and Individuals in Online Community
Committee Chair: Julia M. Rayz

Research in the last decade has prioritized the effects of online texts and online behaviors on user information prediction. However, the previous research overlooks the overall meaning of online texts and more detailed features about users' online behaviors. The purpose of the research is to detect the adopted ideas, the popularity of ideas, and the popularity of individuals by identifying the overall meaning of online texts and the centrality features based on user's online interactions within an online community.

To gain insights into the research questions, the online discussions on *MyStarbucksIdea* website is examined in this research. *MyStarbucksIdea* had launched since 2008 that encouraged people to submit new ideas for improving Starbuck's products and services. Starbucks had adopted hundreds of ideas from this crowdsourcing platform. Based on the example of the *MyStarbucksIdea* community, a new document representation approach, Doc2Vec, synthesized with the users' centrality features was unitized in this research. Additionally, it also is essential to study the surface-level features of online texts, the sentiment features of online texts, and the features of users' online behaviors to determine the idea adoption as well as the popularity of ideas and individuals in the online community. Furthermore, supervised machine learning approaches, including Logistic Regression, Support Vector Machine, and Random Forest, with the adjustments for the imbalanced classes, served as the classifiers for the experiments.

The results of the experiments showed that the classifications of the idea adoption, the popularity of ideas, and the popularity of individuals were all considered successful. The overall meaning of idea texts and user's centrality features were most accurate in detecting the adopted ideas and the popularity of ideas. The overall meaning of idea texts and the features of users' online behaviors were most accurate in detecting the popularity of individuals. These results are in accord with the results of the previous studies, which used behavioral and textual features to predict user information and enhance the previous studies' results by providing the new document embedding approach and the centrality features. The models used in this research can become a much-needed tool for the popularity predictions of future research.

CHAPTER 1. INTRODUCTION

In recent years, considerable attention has been given to the importance of online communities in shaping and spreading public opinion (Cheung & Lee, 2012; Conover et al., 2011; Faraj & Johnson, 2011). User-generated content aggregated from online community members provides information about people's attitudes, opinions, and behaviors. Online platforms not only offer opportunities for people to express their own opinions, but they also accelerate knowledge sharing, influence opinion formation, trigger collective wisdom, and speed up decision-making process (Gruber, 2008). A significant amount of online behavioral and textual data has provided researchers with an excellent opportunity to investigate why certain opinions and individuals become popular in online forums. Surowiecki (2005) stated that the process of aggregating information from people all over the world changes the nature of knowledge production. Lévy (1997) also argued that online environments create a new "knowledge space" that encourages an interactive information flow and increases engagement in civil discussion.

Although the role of online communities in opinion sharing has been explored (Cheung & Lee, 2012; Malhotra & Majchrzak, 2014), there is still little empirical research showing how the overall meaning of online texts combined with various online behaviors make specific ideas and individuals popular in online communities. The techniques used to identify popular ideas and individuals in online communities will be beneficial for companies' word-of-mouth marketing strategies, management information systems, political campaigns, and health interventions, to name a few. To investigate the features that affect the detection of popular ideas and individuals in online communities, this work explores the meaning of user-generated content as well as online interactions involved in the discussions in an online community.

The data for this research comes from the *MyStarbucksIdea* online community. Starbucks, a leading company with a flourishing crowdsourcing community, created the online platform *MyStarbucksIdea* in 2008 to encourage customers to share their experiences, suggestions, and insights on the consumption of Starbucks products, services, and environments (Hossain & Islam, 2015). This co-creation process among customers generated many novel ideas and triggered active discussions on these ideas (Ramaswamy & Gouillart, 2010).

The goal of this work is to investigate the process of detecting adoptable ideas, popular ideas, and popular users in the *MyStarbucksIdea* online community. This was accomplished by using a document embedding approach, Doc2Vec, to represent the overall meaning of all the submitted ideas within the community. We call the representation of the idea as "idea embedding" in this dissertation. Along with idea embedding, online behaviors and surface-level features of the online texts were also studied, to detect idea adoption as well as the popularity of ideas and individuals within the *MyStarbucksIdea* online community. More specifically, this research was undertaken to understand how idea embedding and specific features of online texts and behaviors help to identify adoptable ideas as well as the popularity of ideas and individuals. The models used in this research to detect the adoptable ideas and popularity of ideas and individuals in the online community can become a much-needed tool for future research.

1.1 Scope

This research focuses on detecting the adoptable ideas, the popularity of ideas, and the popularity of individuals through idea embedding, surface-level features of online texts, as well as different features of online behaviors in the *MyStarbucksIdea* online community. Starbucks provided its consumers with an online platform to help improving product development and

service quality. The online platform encouraged customers to generate new ideas for making Starbucks' products, services, and business operations better. Since 2008, Starbucks had adopted several hundred customer ideas from the crowdsourcing platform and developed successful products based on these adopted ideas. Through this platform, Starbucks explored its customers' expectations about improving services and experiences, developing new products, and building customer relationships. This customer co-creation process generates new experiences for customers (Ramaswamy & Gouillart, 2010) and triggers valuable business ideas for the company. By detecting popular ideas and individuals in the *MyStarbucksIdea* community, the tools used in the present study will help to enhance online community flourishing and increase company benefits. Furthermore, the identification of important features of online texts and behaviors studied in the *MyStarbucksIdea* online community will contribute to the detection of salient ideas and individuals in other online platforms for future research.

1.2 Significance

A large body of literature exists on user information prediction by online texts and behaviors. However, the previous literature often overlooks the overall meaning of online texts and comprehensive features based on users' online behaviors. *The purpose of the research presented in this dissertation is to detect the popularity of ideas and individuals by identifying the meaning of online texts and the features of online behaviors that can be used to classify idea adoption as well as the popularity of ideas and individuals within an online community.*

Detecting popular ideas and individuals in online communities will be commercially valuable for companies as they may gain insights into new product development and effective marketing strategy. As previous research shows, opinion leaders in online social blogs usually are

the most popular people in a network and can help product promotion (Li & Du, 2011). The Management Information System benefits from understanding the factors that contribute to helpful online reviews (Huang, Chen, Yen, & Tran, 2015; Mudambi & Schuff, 2010). Identifying popular individuals and ideas is also valuable for political campaigns and health interventions since popular individuals or ideas can attract intensive discussions among stakeholders (Park, 2013), influence other people's opinions (Lam & Schaubroeck, 2000), and possibly change policy making (Nisbet & Kotcher, 2009).

Although online communities have been extensively investigated, thus far there has been relatively little research on the representation of the overall meaning of online texts for popularity predictions. Besides studying the overall meaning of online texts, it also is possible to determine the popularity of individuals and ideas by studying various online behaviors, including commenting, discussing, posting, forwarding, retweeting, and voting. Based on these online behaviors and interactions, we can calculate each user's centrality in the online community. These centrality features can determine the importance and salience of individual users in a community (Chan & Li, 2010; Wasko & Faraj, 2005). Thus, it is essential to consider both textual and behavioral features for the detection of popular ideas and individuals.

The state-of-the-art document representation approach, Doc2Vec, is used to represent user-generated content in the present study. The techniques of Natural Language Processing (NLP) and Machine Learning can process a large number of unstructured texts to identify the essential features for the popularity detection within those texts. Centrality measures help to understand different types of online behaviors and determine user importance as well as user popularity. Many studies have given us useful information on classifying demographic information from online texts, but there is little insight into how certain individuals and ideas become popular online. To

understand this question, the overall meaning of online texts synthesized with detailed behavioral data will be more comprehensive for the popularity detection.

This research attempts to answer the following questions:

- How accurately can *idea adoption* be classified based only on idea embedding?
- How accurately can *idea adoption* be classified based on idea embedding, surface-level features of idea texts, features of online behaviors, and centrality features for the author of the idea?
- How accurately can the *popularity of ideas* be classified based on idea embedding, surface-level features of idea texts, features of online behaviors, and centrality features for the author of the idea?
- How accurately can the *popularity of individuals* be classified based on idea embedding, surface-level features of idea texts, features of online behaviors, and centrality features for the author of the idea?

Answering these questions will contribute to our understanding of the process of popularity detection.

1.3 Assumptions

This dissertation investigates the idea adoption, popularity of individuals, and popularity of ideas in an online community and also identifies significant features of online texts and behaviors. The dissertation aims to understand the different roles these features play in idea adoption as well as the popularity of ideas and individuals. Therefore, the assumptions of this research are:

- There is an association between idea adoption and the overall meaning of idea texts, surface-level features of idea texts, features of online behaviors and centrality features for

the author of the idea. Because of this association, there is a need for social media and online community research to understand online textual and behavioral features, which can be used in future research to predict adoptable ideas within online communities.

- There are hidden relationships between idea popularity and the overall meaning of idea texts, surface-level features of idea texts, sentiment features of online texts, features of online behaviors, and centrality features for the author of the idea. These associations indicate a need to understand online textual and behavioral features that can be used in the future to predict popular ideas in online communities.
- There are associations between individual popularity and the overall meaning of idea texts, surface-level features of idea texts, sentiment features of online texts, features of online behaviors, and centrality features for the author of the idea. These associations indicate the need to understand online textual and behavioral features that can be used in the future to predict popular individuals within online communities.
- Only users that posted ideas can be popular. Popularity in this study is measured based on commenting and is represented as a directed graph. Individual popularity includes only input capture for users who posted ideas on the *MyStarbucksIdea* website.
- There are no offline interactions between users that may affect their ideas or popularity. We can use their online textual and behavioral features to predict idea adoption and popularity for ideas and individuals.
- There is a hidden mechanism in online communities that determines idea adoption and the popularity of ideas and individuals, and it can be detected and predicted.
- Commenting behavior are similar, and thus we can use the data from the *MyStarbucksIdea* website to understand individual popularity in other online communities.

1.4 Limitation of Research

The dataset is based on the actual data in the *MyStarbucksIdea* online community, which results in some limitations for this project. The limitations of this dissertation are:

- Idea adoption, idea popularity, and individual popularity are all highly imbalanced classifications. Only a few ideas out of the many suggestions posted by users become popular and only few of these are ultimately adopted by Starbucks as their new products. Although in the online community everyone can share his or her opinions with others, the individuals who can attract discussion and enhance engagement in the community are scarce. For all these reasons, the dataset in this study is necessarily imbalanced.
- All coffee-related ideas and discussions in the *MyStarbucksIdea* online community were used as the trained data for the classification and detection. There is no new data for the prediction. Since the dataset is highly imbalanced, there is a limited number of positive samples for each question—not enough to be split into classification and prediction datasets. Therefore, the prediction could not be implemented.
- The sparsity of online texts affects the number of meaningful connections among the words, which may influence the performance of the representation of the ideas by Doc2Vec.
- The centrality measure in this research is based on the comments. However, individuals comment at different time points over an interval of years; some comment on an idea within a very short period and others take longer to respond. This is especially true for datasets that span a relatively long time interval, and *MyStarbucksIdea* website was launched over ten years. The difference between infrequent and constant interaction may be significant, but we could not measure this in the present study.

All of these factors result in the limitations of this project.

1.5 Delimitations of Research

Based on the limitations discussed above, the boundaries of this project are as follows:

- No dataset outside of the *MyStarbucksIdea* community was considered.
- Only users who posted a comment or an idea were considered as valid data points in this research.
- Only users who posted at least one idea were considered as valid data points for individual popularity in this research.
- Only ideas related to coffee were considered in this research.
- For the classifications, several resampling methods for balancing the training dataset were utilized to solve the imbalanced class problems in this research.

1.6 Summary

Chapter 1 provides the scope and significance of this dissertation. The assumptions, limitations, and delimitations are also explained in this chapter. Chapter 2 will discuss the relevant literature in developing the research questions for this study.

CHAPTER 2. LITERATURE REVIEW

This research aims to detect the idea adoption, popularity of ideas, and popularity of individuals in the online community. The first section of the literature review outlines the activities in online communities and the associations between online behaviors and semantic content to point out the importance of identifying the features of online texts and behaviors that explain people's online activities. The next section reviews the current research on user information detection related to textual and behavioral features; a review of current machine learning techniques used in user information detection and centrality measures also are included in this section. The third section reviews papers on popularity prediction. The fourth section introduces the latest document representation approach, Doc2Vec, which was used in this research to construct the word embedding of online texts.

2.1 The Interplay between Semantic Content and Online Interactions

Online community members tend to be passionate about discussing the latest issues in which they are interested. Each member plays a different role in an online community. As previous research on commercial online communities has shown, a core group of online community members contributes to the initial development of a product by suggesting insightful ideas, while a peripheral group helps spread and diffuse the latest information (Amrit & Van Hillegersberg, 2010; Crowston, Wei, Li, & Howison, 2006; Fonti & Maoret, 2016; Rullani & Haefliger, 2013; Setia, Rajagopalan, Sambamurthy, & Calantone, 2012). Similar examples abound in the literature.

It is important for online communities to distinguish what kinds of opinions and individuals will be most salient and influential (Fuger, Schimpf, Füller, & Hutter, 2017; Kuppuswamy & Bayus, 2015). However, it is not clear how to define influence and contribution within online

communities. Some researchers think individuals who attract people to discussions are more important than those who just post their ideas because they can increase prosperity in online communities (Füller, Hutter, Hautz, & Matzler, 2014); other researcher shows that the most engaged members eventually will contribute the most economic benefit to the firms (Manchanda, Packard, & Pattabhiramaiah, 2015).

Online communities have a specific nature and mechanism for shaping public opinion and attitude. In the social media age, Facebook, Twitter, and other online platforms offer opportunities for people to express their own opinions. Knowledge sharing on social media has been called a collective wisdom process (Gruber, 2008). It is hoped that the more individuals engage in social networking, the more wisdom they will create. Lévy (1997) argues that the new media environment provides a new "knowledge space" and that it is being transformed by the existing structures of knowledge and power. He argues that new technology promotes online communication, increases civic participation in decision-making, promotes an interactive information flow, and minimizes constraints on communication. Online groups generate collective intelligence and debate meanings and interpretations related to contemporary culture. Furthermore, the information flow is not unidirectional but multidirectional. The information flow is no longer a "two-step flow," where information flows from opinion leaders to the public, but a "multiple-step flow," where everyone can generate the content they share with others. Activities in online communities can be transformed into different features for an online community or social media research. It is worthwhile to investigate how we can translate this online phenomenon into research questions; how we can represent the overall meaning of online texts; and how we can measure online behavior for more detailed analyses.

Research about the association between online behaviors and semantic content has increased noticeably in recent years. Research in cognitive science indicates that there is a positive correlation between the level of social interaction and the similarity of semantic networks among group members (Dugosh & Paulus, 2005). According to Dugosh and Paulus, the semantic network represented human cognition in a group setting. Whether such a conclusion is exaggerated or not, it is clear that there are connections within semantic networks that represent distances between various concepts both individually and in communication between people. A possible interpretation for this is that people tend to hold opinions similar to those with whom they have been interacting.

In the online environment, people interact with each other in ways that are different from daily offline life. Online knowledge sharing is supported by additional mechanisms, such as retweet, forward, and comment (Malhotra & Majchrzak, 2014). However, it could be argued that online interactions are simplified social interactions and that it is easier to quantify them by concentrating only on features that are visible in the online environments. This is not to say that such interactions are entirely simple. For example, people comment on online ideas, and this kind of interaction can trigger discussion and integrate different opinions (Malhotra & Majchrzak, 2014). In this research, we are interested in understanding the interplay between users' online behaviors and online texts to detect popularity. Essentially, this research asks whether individual and idea popularity are correlated to online behavior and self-generated content.

Conover et al. (2011) examined how Twitter facilitates communication between communities with different political attitudes. They identify two network clusters—retweet network and user-to-user mention network—and find clear segregation between communities with different political attitudes in the retweet network. However, this segregation is almost nonexistent

in the user-to-user mentioned network. This result sheds light on the importance of online interactions for connectivity between people of different attitudes.

Rowe and Strohmaier (2014) showed how concepts evolve in online communities by demonstrating the changes within semantic graphs. While little is known about the factors that trigger semantic concept evolution, it is beyond the scope of this dissertation to investigate it. However, it is worth noting that Rowe and Strohmaier suggest that future research should explore the effect of the structure of social networks on semantic content in online communities. Possible reasons for the effects of social networks on semantic content can be uncovered through social network analytics, which extracts features of online interactions (Füller et al., 2014; Füller, Jawecki, & Mühlbacher, 2007) such as commenting, voting, and discussing others' ideas. Semantic content in online communities may be affected over time by these behavioral features, giving prominence to certain textual features (Malhotra & Majchrzak, 2014) and triggering the online mechanism of popular idea formation.

Lewis, Gonzalez, and Kaufman (2012) employed stochastic actor-based modeling to analyze users on Facebook, concluding that people who share similar opinions and attitudes on music and movies more easily become friends. Their findings demonstrate that there is a relationship between online interactions and people's attitudes: people tend to interact with those who share similar opinions online. This also explains the assumption in the dissertation: there is a hidden association between online interactions and the overall meaning of texts. Since online behaviors and online texts may predict and identify each other, this present study aims at elucidating the relationship between them.

The findings described in this section further our understanding of the interplay between the meaning of online texts and user behaviors in online communities.

2.2 User Information Detection from Online Texts and Behaviors

As the rapid development of social media and communication technologies has led to the rise of the information age, much research has made use of the unprecedented scale of online data. In recent years, there has been a dramatic proliferation of research studying the detection of online users' personal information from online texts and behaviors.

2.2.1 User Information Detection from Online Texts

Many studies have analyzed users' posts and tweets by utilizing machine learning techniques to predict users' demographic information, including gender, age, race, and occupation. According to past research, textual features for gender detection can be found in: user tweets (Alowibdi, Buy, & Yu, 2013; Bamman, Eisenstein, & Schnoebelen, 2014; Benton, Mitchell, & Hovy, 2017; Beretta, Maccagnola, Cribbin, & Messina, 2015; Burger, Henderson, Kim, & Zarrella, 2011; Fink, Kopecky, & Morawski, 2012; Liu & Ruths, 2013; Ludu, 2014; Miller, Dickinson, & Hu, 2012; Rao, Yarowsky, Shreevats, & Gupta, 2010; Volkova, Bachrach, Armstrong, & Sharma, 2015); user posts in online platforms other than Twitter (Argamon, Koppel, Pennebaker, & Schler, 2009; Filippova, 2012; Goswami, Sarkar, & Rustagi, 2009; Ikeda, Takamura, & Okumura, 2008; Nowson & Oberlander, 2006; Peersman, Daelemans, & Van Vaerenbergh, 2011; Rao et al., 2011; Reddy, Wellesley, Knight, & del Rey, 2016; Rustagi, Prasath, Goswami, & Sarkar, 2009; Santosh, Joshi, Gupta, & Varma, 2014; Zhang & Zhang, 2010); user names (Alowibdi et al., 2013; Beretta et al., 2015; Burger et al., 2011; Liu & Ruths, 2013; Rao et al., 2011); and user descriptions (Burger et al., 2011). These selected features highly reflect people's characteristics and identities, which is useful for gender prediction.

Work on age detection has focused on: user tweets (Asoh, Ikeda, & Ono, 2012; Beretta et al., 2015; Marquardt et al., 2014; Mechti, Jaoua, Belguith, & Faiz, 2014; Miller et al., 2012; D.

Nguyen, Gravel, Trieschnigg, Meder, & Yeung, 2013; Tuli, 2016; Volkova et al., 2015); user posts (Goswami et al., 2009; Ikeda et al., 2008; D.-P. Nguyen, Gravel, Trieschnigg, & Meder, 2013; D. Nguyen, Smith, & Rosé, 2011; T. Nguyen, Phung, Adams, & Venkatesh, 2011); and user name (Beretta et al., 2015; Siswanto & Khodra, 2013). These selected features reflect people's preferences and experiences that are useful for age prediction. Age usually is viewed as either a simple numeric attribute or as intervals.

2.2.2 Supervised Machine Learning Approaches

The medium described above is used for supervised machine learning classification (and other heuristics) with the following methods: Support Vector Machines (SVM) (Benton et al., 2017; Beretta et al., 2015; Fink et al., 2012; Liu & Ruths, 2013; Peersman et al., 2011; Rao et al., 2010; Santosh et al., 2014; Zhang & Zhang, 2010); Naïve Bayes classification (Alowibdi et al., 2013; Goswami et al., 2009; Miller et al., 2012; Rustagi et al., 2009); logistic regression (Kosinski, Stillwell, & Graepel, 2013; Marquardt et al., 2014; Reddy et al., 2016); Bayesian multinominal regression (Argamon et al., 2009); Bayesian estimation (Asoh et al., 2012); data matching (Asoh et al., 2012; Beretta et al., 2015), Ordinary Least Square (OLS) regression (Culotta, Kumar, & Cutler, 2015); expectation maximization framework (Bamman et al., 2014); and many others..

For other demographic detection, Mohammady and Culotta (2014) studied users' tweets and names through supervised linear regression to predict their race. Based on Facebook users' posts and names, Rao et al. (2011) detected their possible attributes, including gender and ethnicity, by using hierarchical Bayesian models. In addition to many studies on age, gender, and race detection, there is also some research on other demographic detection, such as occupation classifications (Preoţiuc-Pietro, Lampos, & Aletras, 2015; Santosh et al., 2014; Siswanto & Khodra, 2013).

Rosenthal and McKeown focused on lexical features, including writing stylistic features, n-gram features, part-of-speech and collocation features and so on to successfully predict Livejournal users' age, gender, and religion through their posts using supervised machine learning techniques (Rosenthal & McKeown, 2011, 2016). Rustagi et al. (2009) analyzed the stylistic variation of the posts on various blogging platforms to infer users' ages and genders by supervised Naïve Bayes. Siswanto and Khodra (2013) analyzed emoticons used in tweets to predict users' age by SVM.

Online textual features are getting considerable attention, not only for detecting users' demographic information, but also for identifying other personal characteristics, including personality (Golbeck, Robles, Edmondson, & Turner, 2011; Golbeck, Robles, & Turner, 2011; Plank & Hovy, 2015); political orientation (David et al., 2016; Malouf & Mullen, 2008; Pennacchiotti & Popescu, 2011); mental health conditions (Benton et al., 2017; Coppersmith, Dredze, Harman, & Hollingshead, 2015; Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015); and even suicide attempts (Coppersmith, Ngo, Leary, & Wood, 2016; Pestian et al., 2012). Cesare, Grant, and Nsoesie (2017) raise instructional issues related to the classification of user demographics on social media. They insisted that there is a need to develop metadata with possible identity attributes for measuring personal attributes that are difficult to classify.

The above two sections explain how demographic information is predicted using textual features and supervised machine learning methods. The next section will further explain how these predictions are made based on behavioral features.

2.2.3 User Information Detection from Online Behaviors

The accessibility and anonymity of social media and online forums make people feel free to share their thoughts online, which results in an abundance of online texts that contain implicit

information about users' characteristics. Evidence in the above literature has established the practice of predicting personal traits through written text. However, comparatively little research has focused on the relationship between different kinds of online behaviors and online texts.

There have been several studies attempting to identify different social and behavioral types within online communities by determining how users build their social interactions within online communities. What is of interest to us is research based on the frequency of posts. Füller et al. (2007) classified three different user types in an online basketball consumer community based on posting frequency: lurkers, who passively observe others' communications and have no contributions; posters, who contribute to the topics they are interested in; frequent posters, who contribute almost daily to the online communities.

Besides behavioral patterns, social network perspectives can capture the structure of directed (which emphasizes the direction of commenting) and indirect (which does not emphasize the direction of commenting) interactions within communities and further elaborate on user behaviors (Füller et al., 2014). Many of the papers described here use centrality approaches and metrics. It is not unreasonable to suppose that centrality measurements are useful in understanding the popularity of ideas and individuals in online communities.

2.2.4 Centrality Measures Based on Online Interactions

Centrality measurements in this dissertation are based on commenting behaviors, commenting defined as a directed interaction. For instance, A commenting on B's idea is different from B commenting on A's idea based on who initiates a conversation, although both have the commenting frequency of 1. The users in the community are connected through their interactions, which are represented by an edge between two users. The relationships between users are directional, so the direction of an edge indicates who commented and who received the comment.

The definitions of different measures of centrality can be articulated as follows (Kolaczyk & Csárdi, 2014):

1. Degree centrality: Given a vertex V, in a network graph G = (V, E), degree centrality is the count of the number of edges in E incident upon V. In this dissertation, we calculated the number of links a user or idea has with other users. Degree centrality for a directed graph can have two forms (Kolaczyk & Csárdi, 2014):

   1.1 Indegree centrality[1]: The number of edges that a vertex has from other vertices.

   1.2 Outdegree centrality[2]: The number of edges that a vertex has sent to other vertices.

2. Closeness centrality: The average length of the shortest path between a given vertex and all other vertices in the graph. The higher the closeness centrality is, the closer all other vertices are in the network. The measurement is based on the sum of the geodesic distances from each vertex to all the others (Kolaczyk & Csárdi, 2014).

3. Betweenness centrality: The extent to which a vertex is linked to other unconnected vertices. Betweenness centrality quantifies vertices that act as a bridge along the shortest path between two other (groups of) vertices (Kolaczyk & Csárdi, 2014).

4. Eigenvector centrality: For a given graph G= (V, E) with |V| a number of vertices let A = (av, t) be the adjacency matrix if vertex V is linked to vertex T, and av, t = 0. According to this definition, eigenvector centrality is a measure of the influence of a vertex based on how many connections it links to high-scoring vertices. (Kolaczyk & Csárdi, 2014).

[1] The number of comments a user/idea receives. The calculations can base on each user or each idea.
[2] The number of comments a user sent out to other users/ideas.

5. Harmonic centrality: Reversing the sum and reciprocity in the definition of closeness centrality (Rochat, 2009).

6. Eccentricity centrality: The distance between a node and the most distant node based on reciprocal of the maximum of shortest paths in the graph (Jalili et al., 2015; Jalili et al., 2016; Watts & Strogatz, 1998).
7. Clustering coefficient: Measuring to what extent a single node's neighborhood is completed (Watts & Strogatz, 1998).

8. Component number: The number of connected users related to a distinct node in the graph (Hopcroft & Tarjan, 1973; Tarjan, 1972).

Some researches adopted the centrality measures to describe the users' behaviors in the online community. Toral, Martínez-Torres, and Barrero (2010) used the social network approach to identify brokers, defined by betweenness centrality, who act as intermediaries between experts and peripheral users those are identified by degree centrality. They bridge the gap between user types by highlighting information flow and knowledge sharing approaches to engage Open Source Software projects in a co-learning experience within their user communities (Toral et al., 2010). Cross, Laseter, Parker, and Velasquez (2006) classified central connectors, brokers, and peripheral players in a virtual community by utilizing degree centrality and betweenness centrality. Nolker and Zhou (2005) use centrality and behavioral-based measures to identify leaders, motivators, and chatters as the three key member roles in online knowledge-sharing communities. As for the motivations behind online community members, Faraj and Johnson (2011) demonstrated multiple motivations from directed and indirect reciprocity that can co-exist within online communities. Moreover, they demonstrated the existence of community-specific social processes and social norms regulating participation dynamics. Based on the above, we know that user's online

behaviors can be described by centrality measures. However, these centrality features are under-researched and under-discussed regarding their usefulness in supervised machine learning techniques.

2.2.5 Combination of Textual and Behavioral Features

In recent years, growing numbers of research studies have combined both textual and behavioral features to detect user demographics. Culotta et al. (2015) used "follow" relationships to predict Twitter users' gender, race, and ethnicity by using supervised ordinary least squares (OLS) regression. Kosinski et al. (2013) focused on user "likes" to predict Facebook users' sexual orientation, demographic information, religious and political attitudes, personality traits, intelligence, and happiness through linear and logistic regression. Ludu (2014) analyzed user tweets, along with the celebrities they follow, to predict Twitter users' gender, classified by SVM. Pennacchiotti and Popescu (2011) examined user tweets along with user names, photos, dates of creation and friends/followers to predict Twitter users' race by the Gradient boosted decision tree. Past research also analyzed the content of posts as well as some behavioral features – in addition to text features described in the previous section – including the number of friends, posts, and comments a user has to predict *Livejournal* users' ages by supervised linear regression (Rosenthal, 2014; Rosenthal & McKeown, 2011).

However, there is a paucity of thoughtful approaches dealing with online behavior detections from online texts. Thus, this study is an attempt to supplement the findings of studies discussed above by utilizing advanced document embedding approach and centrality measures.

2.3 Popularity Detection from Online Texts and Behaviors

This dissertation shows the extent to which popular ideas and individuals can be predicted by studying online behaviors from online texts. This section describes these predictions by using supervised machine learning methods.

Some companies and organizations have launched online communities to gain collective wisdom and meet needs and inspirations (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). However, very little is known about what makes ideas popular in these communities. In this dissertation, we view online communities that have engaged diverse users to propose different ideas and integrate different opinions.

According to past research, new product reviews are usually diffused online by "heavy users" and "central" users in the network who are defined by their centrality (Iyengar, Van den Bulte, & Valente, 2011). However, what has not been explored are the features of the popular ideas. The question of interest for this dissertation is whether certain features characterize popular ideas.

Some research in the field of Management of Information system has been done to identify the features resulting in popular online reviews. Ghose and Ipeirotis (2011) estimated product review helpfulness by using reviewer characteristics, reviewer history, review readability and review subjectivity by using Random Forrest as the classifier. The results accurately predicted product sales and review helpfulness. Duan, Cao, and Gan (2010) used logistic regression to discover semantic features and successfully predict review helpfulness. Ngo-Ye and Sinha (2012) used the text regression model with dimension reduction techniques to predict review helpfulness. Chen, Qi, and Wang (2012) used polarity features with Conditional Random Fields (CRFs) model to successfully predict review elements. Previous research used IMDB review sentiment and review quality with supervising machine learning approaches to predict movie sales(Yao & Chen, 2013; Yu, Liu, Huang, & An, 2012). Feng and Lin (2016) used Recurrent Neural Networks (RNN)

to predict the ratings of food reviews. Jin et al. (2016) combined the features of review content and user behavior to predict JuiceDB and TripAdvisor review ratings. They first used RNN to learn latent vector representations of the online review, replacing the missing values of aspect ratings with users' reviewing behaviors, and then proposed an optimization framework to predict the review rating.

Based on the research reviewed above, the purpose of this present study was to ascertain the various textual and behavioral features associated with the popularity of certain ideas and individuals in the online community. This study aimed at detecting individual popularity by analyzing online users' written texts to provide new empirical evidence of predicting people's popularity from texts. A document embedding approach, Doc2Vec, will be introduced in the following section.

2.4 Distributed Representations of Document

For text classification questions, the textual input is required to be a fix-length vector to represent the document. The most popular way of constructing a fix-length vector is the bag of words approach (Harris, 1954). However, according to Le and Mikolov (2014), the drawbacks of this approach is (a) the word order is not accounted for, which results in different sentences having the same representation if the same words are used; (b) semantics and distance between words are ignored. To address the problems above, Paragraph Vector, known as Doc2Vec (Document to Vector), is introduced to improve the original text representation approach by carrying the meanings of words into vector space.

Doc2Vec is the improved algorithm of Word2Vec, which is the algorithm used to compute continuous word representation using a two-layer neural network to generate word vectors based

on contexts and overall meaning of words (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Karafiát, Burget, Cernocký, & Khudanpur, 2010; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Yih, & Zweig, 2013). The theoretical origin of Word2Vec is the idea in Linguistics of the distributed hypothesis: "a word is characterized by the company it keeps" (Harris, 1954).

Figure 2.1 shows the framework of Word2Vec. The input words are “the,” “cat,” and “sat,” which are assigned to the word matrix and used to predict the word “on.” The words with similar meanings will be grouped in the close vector space (Le & Mikolov, 2014; Mikolov, Chen, et al., 2013).

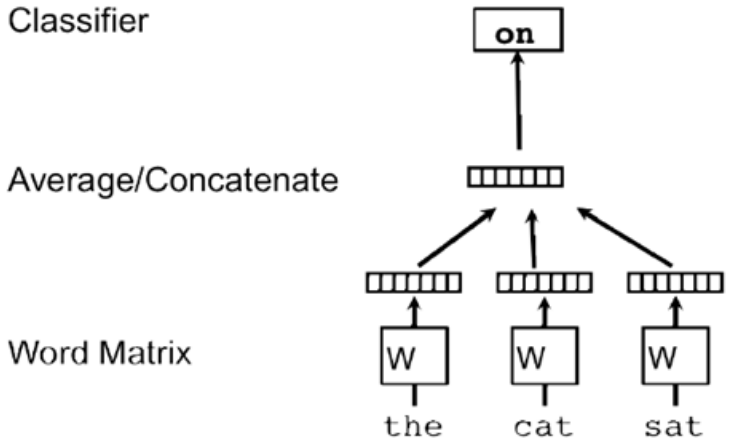

Figure 2.1 A Framework for Word2Vec (Le & Mikolov, 2014)

After the success of Word2Vec, the researchers expanded the word vector to the sentence, paragraph, and even document levels. As shown in Figure 2.2, in the Doc2Vec framework, each document is tagged to a unique vector represented in the word matrix, and each word is also mapped to a vector that also is represented as a matrix. Then, the researchers combined the two matrixes: paragraph vectors as well as input words “the,” “cat,” and “sat” were used to predict the

next word “on” in the given contexts, the framework as shown in Figure 2.2. In this approach, the order of words is preserved in the sentence token. Thus, the model is called the distributed memory model of paragraph vectors (PV-DM).

Doc2Vec also can predict the random word in a given context while ignoring the order of words. This approach is called the distributed bag of words of paragraph vector (PV-DBOW).

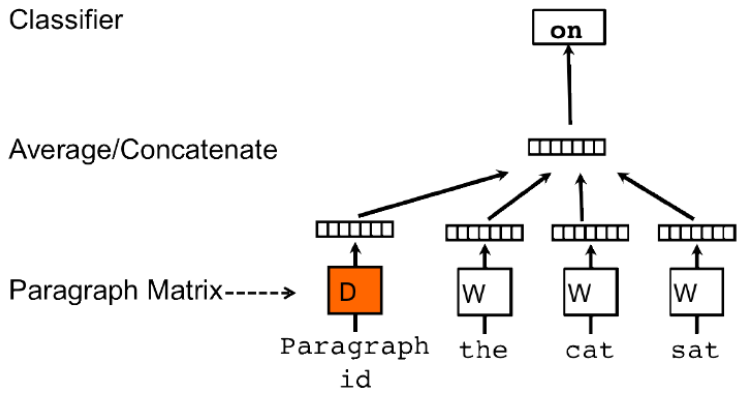

Figure 2.2 A Framework for Doc2Vec-PV-DM (Le & Mikolov, 2014)

As shown in Figure 2.3, the PV-DBOW model concatenates the paragraph vector and word vector to predict the word in a text window without keeping word order. A text window is resampled in each iteration of gradient descent, and we can predict the random word from it.

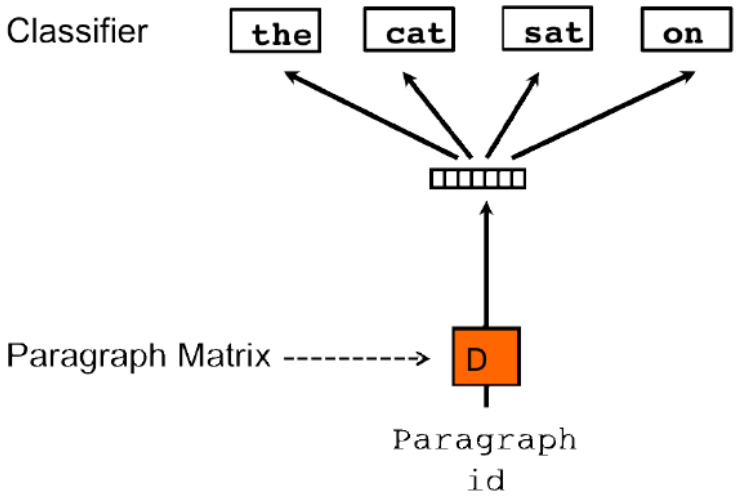

Figure 2.3 A Framework for Doc2Vec PV-DBOW (Le & Mikolov, 2014)

Le and Mikolov (2014) compared the PV-DM and PV-DBOW models by conducting sentiment analysis with IMDB dataset, and the results showed that the PV-DM model outperformed the PV-DBOW model. Lau and Baldwin (2016) also found that the PV-DM model performed better than the PV-DBOW model on the semantic similarity task.

Le and Mikolov (2014) found that Doc2Vec's performance was better than the Recursive Neural Network (RNN), Matrix-Vector-RNN, and Recursive Neural Tensor networks for text classifications. Other research also found Doc2Vec's performance to be superior to the bag of words approaches (Gómez-Adorno et al., 2016; Lau & Baldwin, 2016; Markov, Gómez-Adorno, Posadas-Durán, Sidorov, & Gelbukh, 2016; Niu, Dai, Zhang, & Chen, 2015).

2.5 Imbalanced Class Problem

As we mentioned in Chapter 1, since the only small number of ideas become popular in a community, idea adoption, idea popularity, and individual popularity all have an imbalanced class problem. The ideas adopted by Starbucks are also scarce. Individuals who can gain attention and

become opinion leaders in the online community are also relatively few. So the classifications in this research are all imbalanced classifications.

Imbalanced datasets have been shown to have negative impacts on the performance of classification; thus, resampling methods were used to mitigate the problem (Tang, Zhang, Chawla, & Krasser, 2009). According to Batista, Prati, and Monard (2004), two combinations of resampling methods, SMOTe + Tomek Links and SMOTe + ENN, have good performances for adjusting classifications with a small number of positive samples. SMOTe stands for Synthetic Minority Over-sampling Technique. A pair of samples is referred to as Tomek links if they belong to different classes and are each other's nearest neighbors (Tomek, 1976). Besides the two combination methods, according to the same source (Batista et al., 2004), the over-sampling approach worked better than the under-sampling approach, and random oversampling was better than other over-sampling methods. This research will examine these methods to see which one will produce better results for the *MyStarbucksIdea* dataset.

We used the following imbalanced class adjustment methods reported by (Batista et al., 2004) to adjust the Starbucks dataset of coffee related ideas.

1. Random over-sampling: randomly replicate samples in minority classes
2. Random under-sampling: randomly remove samples in the frequent class
3. SMOTe + Tomek Links: the combination of the Synthetic Minority Over-sampling Technique (Smote) with Tomek links. This approach first over-samples the minority class and then uses Tomek Links as a data cleaning method to better define the class cluster
4. SMOTe + ENN: the combination of Synthetic Minority Over-sampling Technique (SMOTeSMOTe) with Edited Nearest Neighbor (ENN). This approach is first to over-

sample the minority class and then use Edited Nearest Neighbor (ENN) as a data cleaning method to better define the class cluster. In comparison to Tomek Links, ENN removes more samples from the class.

2.6 Summary

This chapter has presented a summary of relevant research on the detection of user information and popular online reviews from online texts and behaviors. Table 2.1 summarizes the textual features and behavioral features used in previous research and outlines existing research gaps.

Table 2.1 The Summary of Textual and Behavioral Features in Literature Review

Category	**Feature**	**Research Gap**
Textual Features	Emoticons	Lack of the representation of the overall meaning of entire documents
	Acronyms	
	Internet slang	
	Collocations	
	Punctuation	
	Capitalization	
	Review length	
	Part-of-Speech Tag	
	N-gram	
	Polarity	
	Subjectivity	
	User name	
Behavioral Features	# of friends	1. Lack of user's centrality measures based on the online interactions with other users 2. Lack of accounting for user commenting on their own posts
	# of posts	
	# of following	
	# of likes	

Regarding textual features, the research gap that remains unexamined is the question of how to use the novel document representation approach to represent the overall meaning of an entire post and detect the popularity of online ideas and individuals. About behavioral features, research has not yet used centrality features based on users' online interactivities and comments on their own ideas to detect the popularity of online ideas and individuals.

Table 2.2 summaries the machine learning approaches used in the previous research. The most commonly used family of approaches is supervised machine learning, which can help us learn the classification from different features of online behaviors and texts.

Table 2.2 The Summary of Machine Learning Approaches in Literature Review

Category	**Models**
Machine Learning Approach	Support Vector Machine (SVM)
	Naïve Bayes Classification
	Logistic Regression
	Bayesian Multinominal Regression
	Bayesian Estimation
	Ordinary Least Square Regression
	Expectation maximization framework
	Convolutional Neural Networks
	Recurrent Neural Network

This dissertation incorporated supervised machine learning techniques with idea embedding and centrality measures to detect idea adoption and popularity of ideas as well as individuals in the online community.

CHAPTER 3. METHODOLOGY

The goal of this research is to gain insights from online texts and behaviors on the classification of the idea adoption, the popularity of ideas, and the popularity of individuals in the Starbucks online community. To achieve this goal, supervised machine learning approaches are used, and four central questions are addressed:

- How accurately can *idea adoption* be classified based only on idea embedding?
- How accurately can *idea adoption* be classified based on idea embedding, surface-level features of idea texts, sentiment features of online texts, features of online behaviors, and centrality features for the author of the idea?
- How accurately can *popularity of ideas* be classified based on idea embedding, surface-level features of idea texts, sentiment features of online texts, features of online behaviors, and centrality features for the author of the idea?
- How accurately can *popularity of individuals* be classified based on idea embedding, surface-level features of idea texts, sentiment features of online texts, features of online behaviors, and centrality features for the author of the idea?

This chapter is divided into five parts. The first part describes the composition of the data set and the method of data collection. The second part provides descriptive statistics related to the dataset itself. The third part proposes the latest document representation approach, Doc2Vec, for analyzing the idea embedding of idea texts. The fourth part centers on methods of feature extraction from online texts and behaviors, and the fifth part explains how to determine whether the descriptive statistics of the extracted features can be used for the classifications.

3.1 Dataset: *MyStarbucksIdeas* Online Platform

As mentioned in the Introduction chapter, *MyStarbucksIdea* is a website where the users can share their ideas about Starbucks products. They also can vote, comment on, and discuss others' ideas submitted on this website.

It is hypothesized that certain types of online behaviors have an impact on the popularity of ideas and individuals. Additionally, some features of online texts and behaviors have a correlation with idea adoption and idea popularity as well as individual popularity.

3.1.1 Data Collection

Every idea submitted on the *MyStarbucksIdea* website has a discussion thread, which fully displays the discussions about each idea (as shown in Figure 3.1). The submitted idea is displayed at the top of the thread, together with the author and the date on which it was posted. Users can vote on the idea or comment. The comments follow the submitted idea. In the example demonstrated in Figure 3.1, four users comment on a submitted idea titled "More sugar-free syrups."

Data from the website was captured on December 12, 2016. The data was stored in three tables of a relational database. The tables are described below and illustrated in Tables 3.1-3:

1. The USER table contains user ID, user name, user location, date of membership, whether the ideas were adopted by Starbucks, and whether the user was selected as a top commenter by Starbucks;
2. The IDEAS table contains user ID, voting scores by other users, when the idea was posted, idea ID, idea title, and idea semantic content;
3. The COMMENTS table contains commenter ID, when the comment was posted, the ID of each comment for an idea, and the content of the comment.

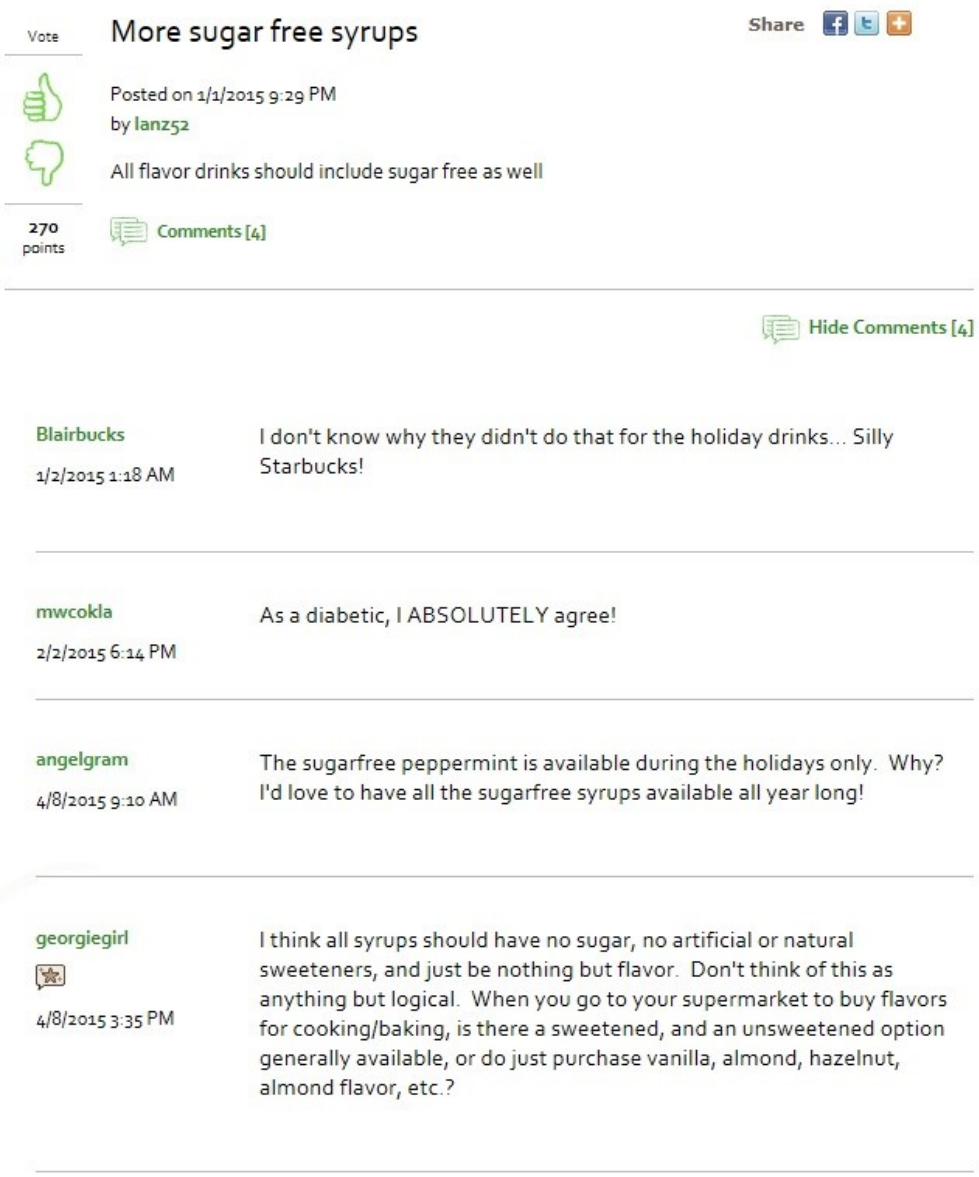

Figure 3.1 Discussion Thread of an Idea

3.1.2 The Composition of the Dataset

As Table 3.1 shows, user information was captured by the *MyStarbucksIdea* website. The user ID is a unique identifier that Starbucks assigned to each user when they registered in the community. User information can be extracted by a unique uniform resource locator (URL) assigned to that user. Every user has a user profile on the website (as seen in Figure 3.2).

Table 3.1 User Information in Discussion Thread

Variable Name	Type/Format	Description
User ID	VARCHAR	The identifier of a user
User Name	VARCHAR	The user's self-reported name
User Location	VARCHAR	The user's self-reported location
User Membership	DATETIME	How long a user had been in the community
Idea Adoption	INTEGER	The number of user's ideas adopted by Starbucks
Top Commenter	BINARY	Elect as a top commenter by Starbucks

Each user can choose a user name to represent themselves in the community. Users do not need to specify personal information such as name and location. The user membership, idea adoption, and top commenter fields were updated by Starbucks.

As Table 3.2 shows, idea information was captured from the *MyStarbucksIdea* discussion thread. The Idea ID is a unique identifier that Starbucks assigns to each idea. Idea information can

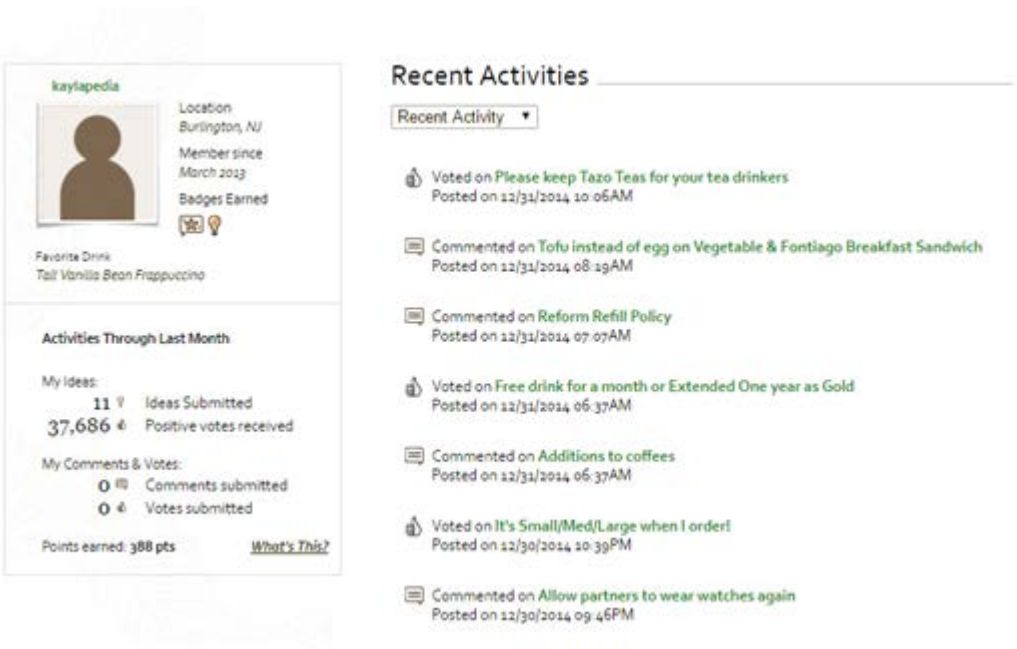

Figure 3.2 The User Profile Page

be extracted by a unique uniform resource locator (URL) assigned to that idea. The voting score was measured by the frequencies of like or dislike votes on an idea by other users.

Table 3.2 Idea Information in Discussion Thread

Variable Name	**Type/Format**	**Description**
User ID	VARCHAR	The identifier of a user who posted the idea
Voting Score	INTEGER	The voting score of the idea
Idea Post Time	DATETIME	The time that the idea was posted
Idea ID	VARCHAR	The identifier for the idea
Idea Title	VARCHAR	The title of the idea
Idea Content	VARCHAR	The semantic content of the idea

As Table 3.3 shows, information about the comments submitted for each idea was also captured.

Table 3.3 Comment Information in Discussion Thread

Variable Name	**Type/Format**	**Description**
Idea ID	VARCHAR	The identifier for the idea
User ID	VARCHAR	The identifier of a user who posted the comment
Comment ID	VARCHAR	The ID of each comment for the idea
Commenting Time	DATETIME	When the comment was posted
Comment Content	VARCHAR	The semantic content of the comment

The user ID is the ID for the user who comments on the idea. The user ID references information in the USER table and the Idea ID reference information in the IDEAS table. The Comment ID is the unique ID of each comment related to the idea in a discussion thread. Each comment has its comment time and content.

3.2 Descriptive Statistics of the Dataset

For this study, we are interested only in coffee-related discussions. There are 15,587 users in total, 9,498 coffee ideas, and 24,533 comments on all coffee ideas. The descriptive statistics of the dataset are summarized in Table 3.4 and Table 3.5. As shown in Table 3.4, there are 9,498 coffee ideas submitted by 8,836 users.

Table 3.4 The Descriptive Statistics of Idea Submission by Users

	Ideas Submission
Mean	1.07
Standard Error	0.0058
Median	1
Mode	1
Standard Deviation	0.486
Sample Variance	0.237
Kurtosis	546.908
Skewness	19.036
Range (min – max) for all Users	0-20
Adopted ideas	436
Count of Ideas	9498
Count of Users who submitting Idea	8836
Count of all Users	15,587

The range for idea submission from a single user is 0 to 20: some users submitted no coffee ideas while others submitted as many as 20. The average number of idea submission by a user is 1.07 (M = 1.07; SD = .486).

Among these submitted ideas, some ideas were adopted when Starbucks decided to use them in future products. Out of 9,498 ideas, 436 have been adopted. The idea adoption is classified as a binary classification. The idea adoption rate is 4.6%.

As shown in Table 3.5, 24,533 comments about coffee ideas have been submitted by 7,822 users, with some users not writing any comments and some writing as many as 845 about coffee ideas. The average number of comment submission by a user is 3.136 (M = 3.136; SD = 23.525).

Table 3.5 The Descriptive Statistics of Comment Submission by Users

	Comments Submission
Mean	3.136
Standard Error	0.266
Median	1
Mode	1
Standard Deviation	23.525
Sample Variance	553.445
Kurtosis	504.634
Skewness	20.617
Range (min – max) for all Users	0-845
Count of Comments	24,533
Count of Users submitting Comment	7,822
Count of all Users	15,587

Table 3.6 shows that out of 9,498 ideas captured, some ideas received no comments, while others received as many as 240. The average number of the comments number for an idea is 2.587 (M = 2.587; SD = .065). To produce a voting score for an idea, each idea can be voted on with a

like or dislike button. The average number of the voting score for each idea is 145,648 (*M* = 145,648; *SD* = .065). The maximum voting score for an idea is 96,120, and the minimum voting scores for an idea is -400. It should be noted that the standard deviation and sample variance of voting scores both are extremely large. The spread of data-point distribution is broad regarding the voting scores of ideas.

Table 3.6 The Descriptive Statistics of Comment Number and Voting Scores of Ideas

	Comment Number	**Voting Score**
Mean	2.587	145.648
Standard Error	0.065	20.725
Median	1	10
Mode	0	10
Standard Deviation	6.337	2019.962
Sample Variance	40.159	4080247.893
Kurtosis	380.904	1255.251
Skewness	14.326	32.066
Range (min – max)	240	96520
Minimum	0	-400
Maximum	240	96120
Count of Comment number \Voting Score	24533	1383510
Count of Idea	9498	9498

3.3 Idea Embedding: The Document Representation Approach

The Doc2Vec Distributed Memory (PV-DM) model was used as the document representation method for displaying users' ideas. The Distributed Memory model was selected instead of the Distributed Bag of Words (PV-DBOW) model to retain word order in the sentences. For dimensionality of the feature vectors, 100 dimensions were designed for the ideas because the length of online ideas is shorter compared to other types of documents. The maximum distance

between the current and predicted word within a sentence (window size) was set to 5 for the same reason. The number of epochs was tuned to 600, since, according to Lau and Baldwin (2016), this is the most optimal value for the Doc2Vec trained model. The Gensim library in Python 3.6.5 (Rehurek & Sojka, 2010) was used for training Doc2Vec model.

The process of training the Doc2Vec model included the following steps:

1. Data cleaning: removing all HTML mark-ups, deleting missing values (there were two ideas with titles but no content), and out of all 9,500 scraped ideas, keeping the 9,498 completed ideas.
2. Data pre-processing: The NLTK library in Python 3.6.5 (version 3.3) was used to remove stop words and to tokenize the sentences (Bird, Klein, & Loper, 2009). Stemming was not used so that information from Internet slang and meaningful terms (e.g., Starbucks) could be preserved.
3. Each idea was tagged with a unique document ID. The tagged ideas were randomly trained for the Doc2Vec model.

3.4 Identifying Features from Online Texts and Behaviors

It is hypothesized that there is an implied mechanism of online discussion at work in the *MyStarbucksIdea* website. This research aims to provide a deeper understanding of the association between specific textual and behavioral features and the popularity of ideas and individuals as well as the idea adoption. In addition to idea embedding, the surface-level features of ideas used in this research include:

1. Surface-level features of the idea texts (Rosenthal & McKeown, 2011): including the frequency of emoticons, Internet slang, punctuation, capitalization, and idea length. These features were captured by the Regular Expression library (version 6.2)

in Python 3.6.5. The descriptive statistics of surface-level feature frequencies are presented in Table 3.7.

2. Sentiment features of ideas (Dang et al., 2010; Jurafsky & Martin, 2014; Pang, Lee, & Vaithyanathan, 2002): idea polarity scores and idea subjectivity scores as captured by the Textblob library (version 0.15.1) in Python 3.6.5. The range of polarity scores is between -1 and +1, with -1 representing the most negative score, and +1 representing the most positive score. The range of subjectivity score is from 0 to 1, with 0 representing the least subjective, and 1 representing the most subjective. The descriptive statistics of sentiment scores are presented in Table 3.8.
3. Behavioral features: including the number of ideas that users post (described in Table 3.4), idea posting time, and frequency of comments on the ideas. The descriptive statistics of sentiment features are presented in Table 3.9.

As shown in Table 3.7, the average number of the frequency of capitalization in each idea is 0.707 (M = 0.707; SD = .4.423); the maximum frequency of capitalization is 181, and the minimum frequency of capitalization is 0. The average number of the frequency of emoticons in each idea is 0.037 (M = 0.037; SD = .206); the maximum frequency of emoticons is 5, and the minimum frequency of emoticons is 0. The average number of the frequency of punctuation signs in each idea is 1.738 (M = 1.738; SD = 2.905); the maximum frequency of punctuation signs is 103, and the minimum frequency of punctuation signs is 0. The average number of Internet slang in each idea is 0.019 (M = 0.019; SD = .153); the maximum frequency of Internet slang is 3, and the minimum frequency of Internet slang is 0. The average number of idea length is 306.001 (M =

306.001; *SD* = 301.086), the maximum length of each idea is 3, and the minimum length of each idea is 2.

Table 3.7 The Descriptive Statistics of Surface-level Features of Idea Texts

	Capitalization	**Emoticon**	**Punctuations**	**Slang**	**Length**
Mean	0.707	0.037	1.738	0.019	306.001
Standard Error	0.045	0.002	0.030	0.002	3.089
Median	0.000	0.000	1.000	0.000	230.000
Mode	0.000	0.000	0.000	0.000	49.000
Standard Deviation	4.423	0.206	2.905	0.153	301.086
Sample Variance	19.561	0.042	8.439	0.023	90652.947
Kurtosis	562.711	74.314	289.605	111.591	81.333
Skewness	20.296	7.044	10.571	9.432	5.233
Range	181.000	5.000	103.000	3.000	7738.000
Minimum	0.000	0.000	0.000	0.000	2.000
Maximum	181.000	5.000	103.000	3.000	7740.000
Sum	6717.000	348.000	16503.000	184.000	2906402.000
Count of Idea	9498	9498	9498	9498	9498

Table 3.8 captures characteristics of the polarity, which measures the idea is positive or negative and subjectivity, which measures how subjective the idea is. These characteristics are helpful because past research demonstrated it is related to idea success (Bayus, 2013). As Table 3.8 shows, the average number of the polarity feature is 0.191 (*M* = 0.019; *SD* = .234). The

maximum score of the polarity feature is 1, and the minimum score of polarity feature is 0. For the subjectivity feature, the average number of the subjectivity feature is 0.476 (*M* = 0.476; *SD* = .225) the maximum score of subjectivity feature is 1, and the minimum score of subjectivity is 0.

Table 3.8 The Descriptive Statistics of Sentiment Feature Scores

	Polarity	**Subjectivity**
Mean	0.191	0.476
Standard Error	0.002	0.002
Median	0.175	0.500
Mode	0.000	0.000
Standard Deviation	0.234	0.225
Sample Variance	0.055	0.050
Kurtosis	1.735	0.207
Skewness	0.310	-0.438
Range	2.000	1.000
Minimum	-1.000	0.000
Maximum	1.000	1.000
Count of Idea	9498	9498

Table 3.9 shows the descriptive statistics of behavior features. As shown in Table 3.9, the average frequency of commenting on their own ideas by each user is 0.132 (*M* = 0.132; *SD* = 0.827). The maximum frequency of commenting on their own ideas is 31, and the minimum frequency of commenting on their own ideas is 0.

Table 3.9 The Descriptive Statistics of Behavior Features

	The frequency of commenting on own ideas
Mean	0.132
Standard Error	0.008
Median	0
Mode	0
Standard Deviation	0.827
Sample Variance	0.684
Kurtosis	533.268
Skewness	18.883
Range	31
Minimum	0
Maximum	31
Sum	1262
Count of Idea	9498

3.5 Identifying Centrality Features from Online Interactions

In addition to elucidating the behavioral features listed above, we also modeled users' centrality in the community based on their online interactions with other users.

To calculate different centrality measures, the first step was to calculate the adjacency matrix. The adjacency matrix has columns representing users who wrote an idea and rows representing users who commented on an idea of a user in that column. We acquired the 8,836*7,822 directional matrix, representing the input of 8,836 users who submitted the ideas and 7,822 users who commented. Based on this information, we visualized the user-to-user interaction network as the directed graph (see Figure 3.3). R (version 3.5.0) and Gephi (version 0.9.1) were used for the social network analysis and visualization of the completed user-to-user interaction network characterizing the interactions in *MyStarbucksIdea*. As seen in Figure 3.3, the online

interactions in *MyStarbucksIdea* are very active. The graph is a directed graph, calculated from the frequency of commenting. The users in the more central positions in the network are considered to be more important (Luo, 2010) because more people comment on their ideas. This is the proxy of individual popularity in this research.

An adjacency matrix based on user-to-user commenting frequencies was used to calculate degree, in-degree, out-degree, and degree centrality and other centrality measures based on the formulae described in Chapter 2, including eccentricity centrality, closeness centrality, betweenness centrality, harmonic centrality, component number, and eigencentrality centrality. The descriptive statistics of user centralities are presented in Table 3.10 and Table 3.11.

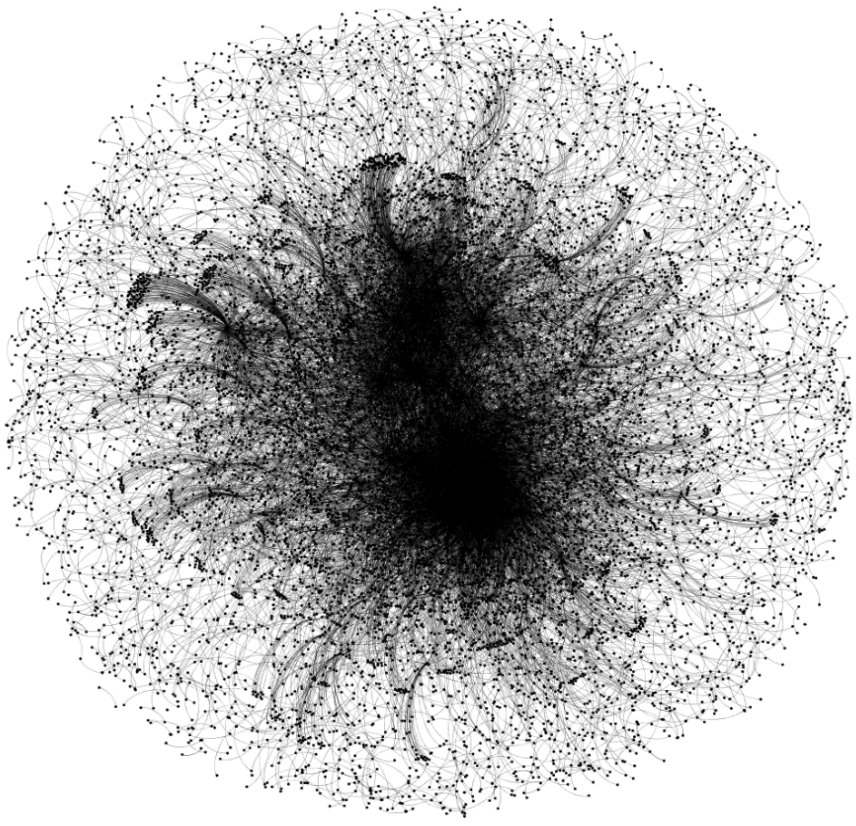

Figure 3.3 The User-to-User Interaction Network of Starbucks Online Community

As shown in Table 3.10, the average number of indegree per user is 4.067 (*M* = 4.067; *SD* = 15.176), with the maximum score at 240, and the minimum score at 0. The average number of outdegree per user is 3.77 (*M* = 3.77; *SD* = 35.296), the maximum score is 845, and the minimum score is 0. The average number of degree per user is 7.837 (*M* = 7.837; *SD* = 47.258), the maximum score is 888, and the minimum score is 0. The average number of eccentricity per user is 0.25 (*M* = 0.25; *SD* = 1.243), the maximum score is 12, and the minimum score is 0. The standard deviations for degree centralities varies significantly, which means that users' online behaviors and popularity levels are extremely different. The average number of closeness centrality per user is 0.068 (*M* = 0.068; *SD* = 0.24), the maximum score is 1, and the minimum score is 0. The average number of eccentricity per user is 0.25 (*M* = 0.25; *SD* = 1.24), the maximum score is 12, and the minimum score is 0.

Table 3.10 The Descriptive Statistics of Centrality Features

	Indegree	**Outdegree**	**Degree**	**Eccentricity**	**Closeness**
Mean	4.067	3.771	7.837	0.250	0.068
Standard Error	0.155	0.362	0.484	0.012	0.002
Median	1	0	1	0	0
Mode	0	0	0	0	0
Standard Deviation	15.175	35.299	47.258	1.243	0.242
Sample Variance	230.309	1245.842	2233.396	1.545	0.058
Kurtosis	111.507	175.236	138.808	42.648	10.173
Skewness	9.861	12.441	11.178	6.435	3.446
Range	240	845	888	12	1
Minimum	0	0	0	0	0
Maximum	240	845	888	12	1
Among the number of comments	24, 533	24,533	24,533	24,533	24,533

As shown in Table 3.11, the average number of harmonic centrality per user is 0.070 (*M* = 0.070; *SD* = 0.245), the maximum score is 1, and the minimum score is 0. The average number of betweenness centrality per user is 1,984.963 (*M* = 1,984.963; *SD* = 18505.38), and the betweenness centrality among different users varies significantly. The maximum betweenness score is 365084, and the minimum betweenness score is 0. The average number of component number for per user is 41.922 (*M* = 41.922, *SD* = 148.4577), the maximum score is 850, and the minimum score is 0.

The average number of clustering per user is 0.039 (*M* = 0.039, *SD* = 0.044), the maximum score is 1, and the minimum score is 0. The average number of eigencentrality for per user is 0.004 (*M* = 0.004, *SD* = 1.243), the maximum score is 1, and the minimum score is 0.

Table 3.11 The Descriptive Statistics of Centrality Features

	Harmonic	**Betweenness**	**Component Number**	**Clustering**	**Eigencentrality**
Mean	0.070	1984.963	41.922	0.039	0.004
Standard Error	0.003	189.881	1.523	0.001	0.000
Median	0.000	0.000	0.000	0.000	0.000
Mode	0.000	0.000	0.000	0.000	0.000
Standard Deviation	0.245	18505.379	148.458	0.108	0.044
Sample Variance	0.060	342449063.600	22039.690	0.012	0.002
Kurtosis	9.766	174.871	13.587	14.966	459.755
Skewness	3.387	12.080	3.776	3.510	20.831
Range	1.000	365084.021	850.000	1.000	1.000
Minimum	0.000	0.000	0.000	0.000	0.000
Maximum	1.000	365084.021	850.000	1.000	1.000
Sum	667.707	18853175.250	398176.000	374.540	36.097
Count	9498.000	9498.000	9498.000	9498.000	9498.000

3.6 Supervised Machine Learning Approaches

To detect the idea adoption, the popularity of ideas, and the popularity of individuals, Scikit-learn (version 0.19.1) in Python 3.6.5 was utilized for three Supervised Machine Learning approaches, including Logistic Regression, Support Vector Machine (SVM), and Random Forrest (Buitinck et al., 2013; Pedregosa et al., 2011). For the adjustments of imbalanced classes, all the resampling techniques were implemented by Imbalanced-learn (version 0.3.3) in Python 3.6.5 (Lemaître, Nogueira, & Aridas, 2017).

3.7 Summary

This dissertation uses various textual and behavioral features (as summarized in Table 3.12) to detect the popularity of online ideas and individuals and to understand the idea adoption better. Even though previous studies combined these features for demographics predictions, research has not yet considered the overall meaning of texts, in combination with centrality features, for popularity prediction. This dissertation is exploratory and tries to find possible associations among online behavioral and textual features, as well as increase understanding of the prediction of idea adoption, popularity of ideas, and popularity of individuals.

Table 3.12 The Summary of Feature Groups

Group	**Feature**	**Description /Examples**
Surface-level features of idea texts	Emoticons	☺
	Internet Slangs	LOL
	Punctuation	!!!
	Capitalization	COOL
	Sentence Length	66
Idea Embedding	The overall meaning of idea texts	Using Doc2Vec PV-DM model to represent the overall meaning of idea texts
Sentiment features	Polarity	How positive the idea is
	Subjectivity	How subjective the idea is
Features of online behaviors	# of ideas	The number of ideas the user post in total
	# of comments	The number of comments the user post in total (as the same with)
	Posting time and day	Idea A is posted at 9 pm on Monday
	Commenting on the idea	Users comment on their submitted idea
Centrality Features	Indegree centrality	The Indegree centrality of the user
	Outdegree centrality	The Outdegree centrality of the user
	Closeness centrality	The Closeness centrality of the user
	Betweenness centrality	The Betweenness centrality of the user
	Eigenvector centrality	The Eigenvector centrality of the user
	Degree centrality	The Degree centrality of the user
	Eccentricity centrality	The Eccentricity centrality of the user
	Harmonic centrality	The Harmonic centrality of the user
	Component Number	The Number of connected users of the user
	Clustering	How completed the neighborhood is for the user who post the idea

CHAPTER 4. RESULTS

Some interesting findings emerged from the experiments in this dissertation. This chapter summarizes results about the detection for the popularity of ideas and individuals in the community as well as idea adoption.

4.1 The Classification of Idea Adoption by Idea Embedding

To predict whether the idea is adoptable or not, the Doc2Vec word embedding representation approach was used for representing the idea texts. Three Supervised Machine Learning approaches, including Logistic Regression, Support Vector Machine (SVM), and Random Forrest were used as classifiers with cross-validation for the evaluation. Once a balanced dataset is achieved, since we were working with a binary classifier, the classification was considered successful if the results were better than chance (50%). The area under ROC curve (AUC), accuracy, precision, recall, and F1 measure were used as the classification evaluation metrics (Tang et al., 2009). The definitions are as follows:

- Precision = $\frac{True\ Positive}{True\ Positive+False\ Positive}$
- Recall = $\frac{True\ positive}{True\ Positive+False\ Negtive}$
- Accuracy = $\frac{True\ Positive\ +\ False\ Positive}{True\ Positive\ +\ False\ Positive\ +\ True\ Negtive\ +\ False\ Negative}$
- F1 Score = $2 * \frac{Precision*Recall}{Precision+Recall}$

4.1.1 Imbalanced Class Problem

The idea adoption was treated as a binary classification problem, based on whether or not ideas were adopted by Starbucks. There were 436 ideas out of 9,498 that were adopted by Starbucks in the coffee group. The dataset was imbalanced as the adoption rate was only around 4.6%, which meant that even if all adopted ideas were incorrectly classified as non-adopted, the accuracy for classifying the non-adopted ideas would still be at 95%.

We used the imbalanced class adjustment methods outlined in the literature review (Batista et al., 2004), including random over-sampling, random under-sampling, SMOTe + Tomek Links, and SMOTe + ENN—to analyze the Starbucks data set of coffee related ideas.

4.1.2 The Evaluation of Resampling Methods

Table 4.1 summarizes the results of idea classification using the Doc2Vec idea embedding approach. The split validation approach is used to train the classifiers (split ratio = .75). The results are presented using five different measures: overall accuracy, overall AUC, overall precision, overall recall, and overall F measure (as shown in the first column of Table 4.1). The Resampled Sample column demonstrates the number of resampled samples: 0 represents the class for non-adopted ideas, 1 represents the class for adopted ideas, and the number followed by the class is the number of resampled samples.

As can be seen in Table 4.1, for Logistic Regression and SVM, SMOTe+ ENN produce the best results, followed by SMOTe+ Tomek Links and Random over-sampling. For the Random Forest classifier, Random over-sampling is most likely to produce the best results. The results are reasonable. Because SMOTe+ Tomek Links removed the samples between classes, this method produced the best results for SVM and Logistic Regression. For Random Forest, more information about the samples will help to produce better results.

Table 4.1 The Summary of Classification Performance for Idea Adoption by Idea Embedding

Measure	Method	Resampled Sample	LR	SVM	RF
Overall Accuracy	Random Over-sampling	(0, 9062) (1, 9062)	0.63	0.56	**0.99**
	Random Under-sampling	(0, 436) (1, 436)	0.5	0.49	0.5
	SMOTe + Tomek Links	(0, 9062) (1, 9062)	0.64	0.58	0.95
	SMOTe + ENN	(0, 3856) (1, 9059)	**0.69**	**0.61**	0.95
Overall AUC	Random Over-sampling	(0, 9062) (1, 9062)	0.68	0.56	**0.99**
	Random Under-sampling	(0, 436) (1, 436)	0.51	0.49	0.53
	SMOTe + Tomek Links	(0, 9062) (1, 9062)	0.66	0.58	**0.99**
	SMOTe + ENN	(0, 3856) (1, 9059)	**0.7**	**0.59**	0.98
Overall Precision	Random Over-sampling	(0, 9062) (1, 9062)	0.63	0.56	**1**
	Random Under-sampling	(0, 436) (1, 436)	0.51	0.49	0.49
	SMOTe + Tomek Links	(0, 9062) (1, 9062)	0.64	0.59	0.95
	SMOTe + ENN	(0, 3856) (1, 9059)	**0.71**	**0.65**	0.95
Overall Recall	Random Over-sampling	(0, 9062) (1, 9062)	0.63	0.56	**1**
	Random Under-sampling	(0, 436) (1, 436)	0.51	0.49	0.5
	SMOTe+ Tomek Links	(0, 9062) (1, 9062)	0.64	0.59	0.95
	SMOTe SMOTe+ ENN	(0, 3856) (1, 9059)	**0.69**	**0.61**	0.95
Overall F1 Score	Random Over-sampling	(0, 9062) (1, 9062)	0.63	0.56	**1**
	Random Under-sampling	(0, 436) (1, 436)	0.51	0.49	0.49
	SMOTe + Tomek Links	(0, 9062) (1, 9062)	0.64	0.58	0.95
	SMOTe+ ENN	(0, 3856) (1, 9059)	**0.7**	**0.62**	0.95

Note: In Resampled Sample column, 0 is the class for non-adopted ideas, 1 is the class for adopted ideas, and the number followed by the class is the number of resampled samples.

Based on the overall results and the corpus, SMOTe+ ENN was selected to use in experiments about idea classification due to its' best performance for Logistic Regression and SVM, as discussed in the next sections.

4.1.3 The Evaluation of Idea Classification

According to Table 4.1, Random Forest is found to have the highest performance on all the evaluation metrics. After the imbalance class adjustment, the performance is over 95% accuracy, AUC, and F-measure for Random Forest Classifier; around 70 % accuracy, AUC, and F-measure for Logistic regression; and 60% accuracy, AUC, and F-measure for SVM. The classification was considered successful because all the results are better than chance (50%).

4.2 The Classification of Idea Adoption by Idea Embedding, Surface-Level Features, Sentiment Features, Behavior Features, and Centrality Features

In the previous section, the Doc2Vec idea embedding approach was used to represent the idea texts and predict whether the idea is adoptable or not. For this experiment, other features were added to the model for idea adoption with the goal of performance improvement. Each idea was characterized by its surface-level features of idea texts, sentiment features, behavioral features, and centrality features (described in Chapter 3). Idea embedding was considered to compare individual and combined performance for the classification.

We represented the number of emoticons, punctuation marks, internet slangs, capitalizations, and idea length as surface-level features; the polarity score and subjectivity scores as sentiment features; the number of idea submission per user, the frequency at which user comment on their own ideas, and idea posting time as behavior features; Indegree, Outdegree, Degree and all centrality measures as centrality features. These four feature groups, in addition to

the idea embedding, then were used as features for classifying each idea. The same Machine Learning classifiers and measures were used for the evaluation.

4.2.1 Individual Feature Selection

In the first model, we separated all the single features for the idea adoption classification, to find the single feature importance by utilizing Random Forest Classifier. Random Forest classifier was selected because it demonstrated the best performance in the dataset. The formula of Random Forest Classifier is:

$$Classification = Bias + feature\_1\ contribution + \cdots + feature\_n\ contribution$$

This research adopted mean decrease impurity method of Random Forest Classifier, where the optimal condition is chosen based on information gain for the feature selection of classification (Archer & Kimes, 2008; Guyon & Elisseeff, 2003; Menze et al., 2009; Peng, Long, & Ding, 2005; Saeys, Abeel, & Van de Peer, 2008).

After the imbalanced class adjustment, the positive contribution to the idea adoption came from Outdegree (removing it decreases model performance by 7.4%); Eigenvector centrality (removing it decreases model performance by 5.2%); Number of submitted Ideas (removing it decreases model performance by 4.5 %); Indegree (removing it decreases model performance by 4.2%); Degree (removing it decreases model performance by 4.1%); Eccentricity centrality (removing it decreases model performance by 4.1%); Betweenness centrality (removing it decreases model performance by 3.2%); and three components of the idea embedding vector, corresponding to the 29^{th}, 100^{th}, and 54^{th} dimension of the idea embedding parameters (removing each of them decreases model performance by 1.0%). The results are shown in Table 4.2.

Table 4.2 The Summary of Feature Importance for Idea Adoption by Random Forest Classifier

Feature	**Contribution**
Outdegree	0.074
Eigenvector centrality	0.052
Number of submitted Ideas	0.045
Indegree	0.042
Degree	0.041
Eccentricity centrality	0.041
Betweenness centrality	0.032
Idea Embedding 29^{th} dimension	0.010
Idea Embedding 100^{th} dimension	0.010
Idea Embedding 54^{th} dimension	0.010

4.2.2 Individual Feature Group

As we mentioned in Section 4.1.2, SMOTe + ENN was chosen to use as the main resampling method for the imbalanced class adjustment, based on its best performance on the dataset. Resampling the dataset resulted in 7,272 samples for the non-adopted ideas and 7,609 for the adopted ideas. The dataset was then split into test and training sets for the experiments.

The results of each group's features (surface-level features of idea texts, sentiment, behavior, and centrality feature groups) are shown below. The performance of the surface-level feature group for idea adoption classification is shown in Table 4.3. The performance of sentiment feature group for idea adoption classification is shown in Table 4.4. The performance of behavior feature group for idea adoption classification is shown in Table 4.5. Finally, the performance of centrality feature group for idea adoption classification is shown in Table 4.6.

As seen in Table 4.3, Random Forest was found to have the highest performance out of all the evaluation metrics for the surface-level feature group. The performance was over 0.99 for accuracy, AUC, and F-measure. Logistic Regression performs at over 80 % accuracy, AUC, and F-measure. For SVM, the performance was over 70% accuracy, AUC, and F-measure. Idea adoption classification by surface-level feature group was considered successful because all the results were better than chance.

Table 4.3 The Summary of Performance for Surface-Level Feature Group on Idea Classification

Measure	Method	LR	SVM	RF
Overall Accuracy	SMOTe + ENN	0.8	0.74	0.99
Overall AUC	SMOTe + ENN	0.85	0.83	0.99
Overall Precision	SMOTe + ENN	0.83	0.87	0.99
Overall Recall	SMOTe + ENN	0.8	0.74	0.99
Overall F-measure	SMOTe + ENN	0.81	0.76	0.99

According to Table 4.4, Random Forest was still found to have the highest performance out of all the evaluation metrics for the sentiment feature group. After the imbalanced class adjustment, Random Forest Classifier performed at over 93% accuracy, AUC, and F-measure. For Logistic regression, the performance was over 50 % accuracy, AUC, and F-measure. For SVM, the performance is over 50% Accuracy, AUC, and F-measure for SVM. Classification by the sentiment feature was considered successful because all the results are better than chance.

Table 4.4 The Summary of Performance for Sentiment Feature on Idea Classification

Measure	Method	LR	SVM	RF
Overall Accuracy (%)	SMOTe+ ENN	0.56	0.53	0.93
Overall AUC (%)	SMOTe+ ENN	0.56	0.54	0.98
Overall Precision (%)	SMOTe+ ENN	0.53	0.53	0.93
Overall Recall (%)	SMOTe+ ENN	0.53	0.53	0.93
Overall F-measure (%)	SMOTe+ ENN	0.53	0.52	0.93

According to Table 4.5, Random Forest was found to have the highest performance out of all the evaluation metrics for the behavior feature group. After the imbalance class adjustment, the performance was over 98% accuracy, AUC, and F-measure. For Logistic regression, the performance was over 70 % accuracy, AUC, and F-measure. SVM performed at over 70% accuracy, AUC, and F-measure. The classification by behavior feature was considered successful because all the results are better than chance.

Table 4.5 The Summary of Performance for Behavior Feature Group on Idea Classification

Measure	Method	LR	SVM	RF
Overall Accuracy (%)	SMOTe+ ENN	0.75	0.71	0.98
Overall AUC (%)	SMOTe+ ENN	0.82	0.89	0.99
Overall Precision (%)	SMOTe+ ENN	0.77	0.79	0.98
Overall Recall (%)	SMOTe+ ENN	0.75	0.71	0.98
Overall F-measure (%)	SMOTe+ ENN	0.74	0.69	0.98

According to Table 4.6, Random Forest was found to have the highest performance out of all the evaluation metrics for the centrality feature group. After the imbalance class adjustment, the performance was over 99% accuracy, AUC, and F-measure. For Logistic regression, the performance was over 70 % accuracy, AUC, and F-measure. SVM performed at over 70% accuracy, AUC, and F-measure. The classification by centrality feature group was considered successful because all the results are better than chance.

Table 4.6 The Summary of Performance for Centrality Feature Group on Idea Classification

Measure	Method	LR	SVM	RF
Overall Accuracy (%)	SMOTe+ ENN	0.72	0.72	0.99
Overall AUC (%)	SMOTe+ ENN	0.86	0.80	0.99
Overall Precision (%)	SMOTe + ENN	0.86	0.86	0.99
Overall Recall (%)	SMOTe + ENN	0.83	0.72	0.99
Overall F-measure (%)	SMOTe + ENN	0.84	0.74	0.99

A comparison of the accuracy of the worst classifier, SVM, for the individual feature groups is shown in Figure 4.1. The SVM classifier was chosen, since its results could potentially be improved by combining several features.

4.2.3 Feature Groups in Combinations

Single features may not be the best choice for the classification, the reasons as shown by Stuart, Tazhibayeva, Wagoner, and Taylor (2013). This section describes the experiments with multiple feature groups. Following the methodology of Stuart et al. (2013), the features were iteratively added to a set with the highest performance until such additions no longer improve the results.

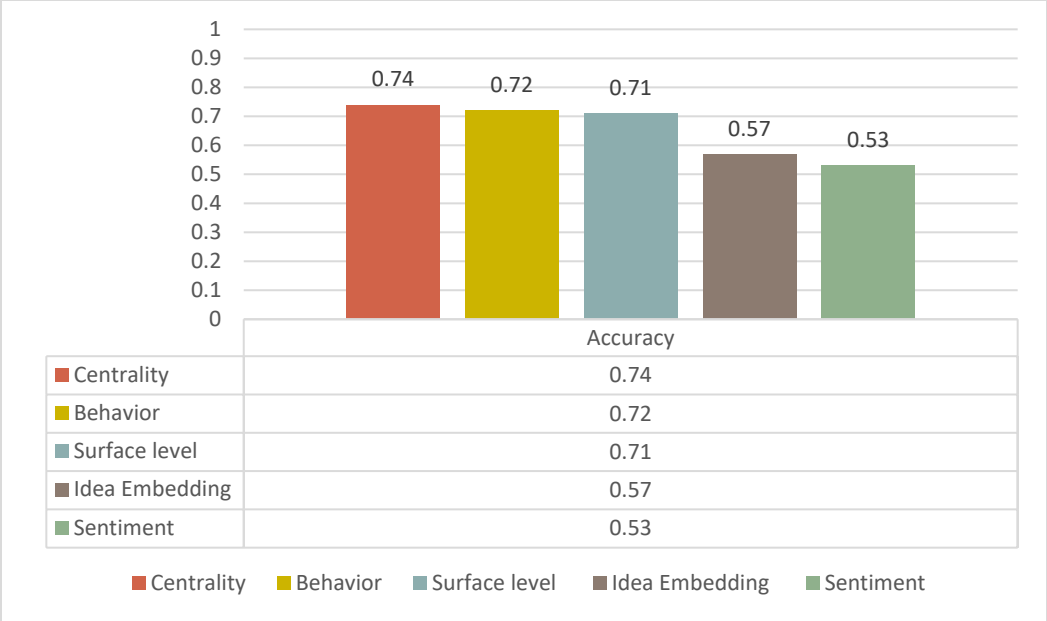

Figure 4.1 The Accuracy of Single Feature Groups for SVM Classifier

According to Figure 4.1, the Top 2 feature groups for the idea adoption classification were the centrality feature group and behavior feature group. Since they were most accurate in classifying the classification of idea adoption. Centrality feature group and behavior feature group combined into a feature set called Top 2 feature group. The other three feature group were added to Top 2 feature group, producing the results of SVM classification, as shown in Figure 4.2.

In the second iteration, the Top 2 feature combination performed the worst. In decreasing order, the accuracy of feature sets was Top 2 + Idea Imbedding (accuracy = 0.75) and Top 2+ Sentiment (accuracy = 0.74), followed by Top + surface-level feature (accuracy = 0.73) and Top 2 feature group by themselves (accuracy = 0.53).

Figure 4.2 The Accuracy of Top2 and Combined Feature Groups for SVM Classifier

For the next iteration, we combined the Top 2 + Idea embedding as a new feature set called Top 3. The other two feature group were added to the Top 3 feature group, resulting in the Top 3 + sentiment feature group and the Top 3 + surface-level feature group. The accuracy of the SVM classification is shown in Figure 4.3.

In this iteration, the accuracy results were very close. The most accurate feature set was Top 3 + Sentiment (accuracy = 0.761), referred to as Top 4. The next most accurate was the Top 3 + surface-level feature group with an accuracy of 0.75. The accuracy of Top 3 remained at 0.75.

In the final iteration, we combined all five features as a new feature set and compared it to Top 4. The accuracy of the 5-feature set for the SVM classifier was 0.765, slightly higher than Top 4 in the last iteration.

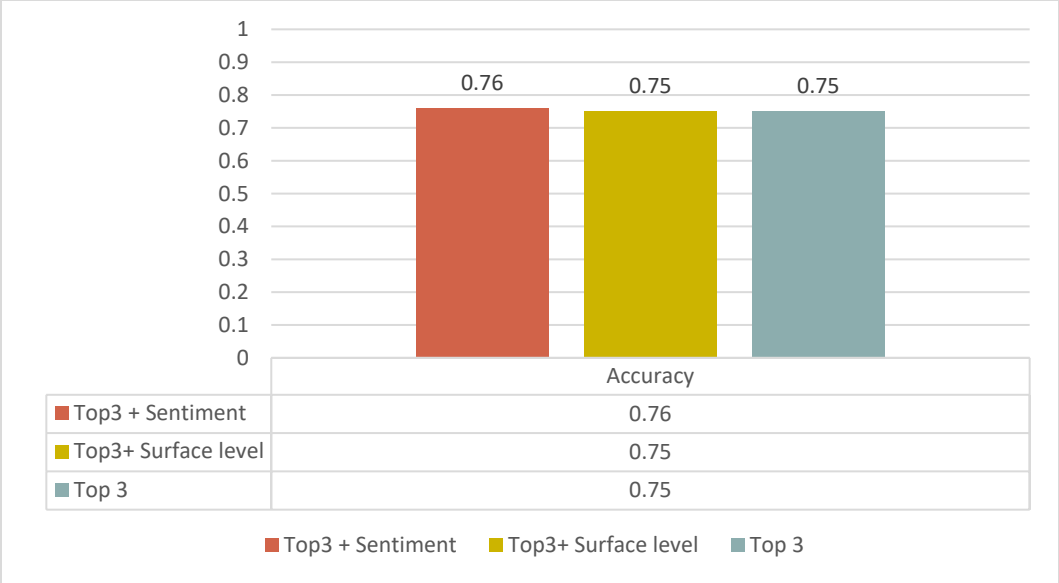

Figure 4.3 The Accuracy of Top3 and Combined Feature for SVM Classifier

The findings indicate that accuracy was increased when we combined more features in the model. These results are in accord with the results of previous studies that have used behavioral and textual features to predict classification and enhance the previous studies' results by providing the latest word embedding approach and centrality features.

4.3 The Classification of Idea Popularity by Idea Embedding, Surface-Level Features, Sentiment Features, Behavior Features, and Centrality Features

In section 4.2, the Doc2Vec word embedding approach and other features of online texts and behaviors were used to predict whether the idea would be adopted or not. In this section, we use the same feature sets to predict idea popularity.

Idea popularity was measured by the number of comments for each idea, and based on this criterion, ideas were divided into two buckets: high popularity and low popularity by the mean of comment number for per idea. To detect the popularity of an idea (eventually resulting in the

number of comments), each idea was characterized with the surface-level feature group, sentiment feature group, behavioral feature group, and centrality feature group listed in chapter 3, as well as with the idea embedding. However, Indegree centrality is associated with idea popularity, so Indegree and degree centrality were removed from the centrality feature group, only used other centrality features for the classification and compared their individual and combined performances.

SVM was used in this experiment with cross-validation for the evaluation. The classification was considered successful if the results were better than chance. Accuracy was used as the classification evaluation metrics (Tang et al., 2009).

4.3.1 Individual Feature Selection

In the individual model, we separated all the single features for the idea popularity regression, to find the single feature importance by utilizing Random Forest regression. The formula of Random Forest Classifier is:

$$Regression = Bias + feature\_1\ contribution + \cdots + feature\_n\ contribution$$

This research adopted mean decrease impurity method of Random Forest regression model, where the optimal condition is chosen based on information gain for the regression model for the feature selection (Archer & Kimes, 2008; Guyon & Elisseeff, 2003; Menze et al., 2009; Peng et al., 2005; Saeys et al., 2008)..

The positive contribution to the idea adoption came from Eigenvector centrality (removing it decreases model performance by 27.6%); comment on own ideas (removing it decreases model performance by 15%); one of the idea embedding components, corresponding to the 49th dimension of the idea embedding parameters (removing it decreases model performance by 5.6 %); (removing it decreases model performance by 4.2%); clustering coefficient (removing it decreases model performance by 3.5%); another one of the idea embedding components, corresponding to

the 15th dimension of the idea embedding parameters (removing it decreases model performance by 3.2%); the number of submitted ideas (removing it decreases model performance by 1.8%); and four components of the idea embedding, corresponding to the 88th, 70th, 9th, 69th dimension of the idea embedding parameters (removing it decreases model performance by from 1.7% to 1.3%).

Table 4.7 The Summary of Feature Importance for Idea Popularity by Random Forest Regression

Feature	**Contribution**
Eigenvector centrality	0.276
Comment on own ideas	0.150
Idea Embedding 49th dimension	0.056
Clustering	0.035
Idea Embedding 15th dimension	0.032
The number of submitted ideas	0.018
Idea Embedding 88th dimension	0.017
Idea Embedding 70th dimension	0.016
Idea Embedding 9th dimension	0.015
Idea Embedding 69th dimension	0.013

4.3.2 Individual Feature Group

In this section, we tested each feature group separately over the ideas to examine which feature group was most important in identifying the popularity of ideas. We used the accuracy of SVM to examine the performance of the feature groups and to compare the performance of the individual feature group instead of the entire feature set.

As seen in Figure 4.4, the top 2 feature groups for identifying the popular ideas are behavior feature group (accuracy = 0.71) and surface-level feature group (accuracy = 0.706), followed by sentiment (accuracy = 0.704), idea embedding (accuracy = 0.543), and centrality features group (accuracy = 0.525).

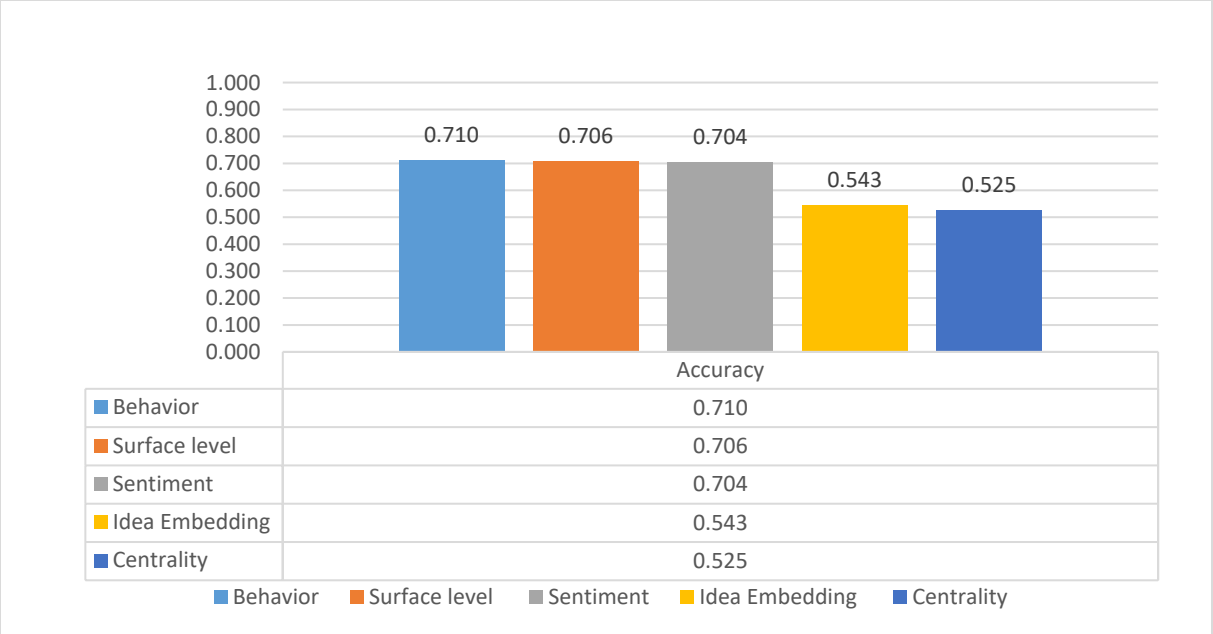

Figure 4.4 The Accuracy of Single Feature Group of SVM for Idea Popularity

4.3.3 Feature Groups in Combinations

After we learned how each feature group performed individually, we focused on determining what combination of feature groups would lead to the best performance for idea popularity classification. From the last section, we know that the Top 2 features for idea classification are behavior feature group and surface-level feature group. They are most accurate in classifying idea popularity.

Therefore, we next combined them into the same feature set, called the Top 2 feature. The other three new feature sets also were built from the Top 2 feature group: Top 2 + sentiment feature group, Top 2 + idea embedding feature group, and Top2 + centrality feature group. Then we re-tested these four new datasets using the SVM classifier to find the most accurate performance.

Figure 4.5 presents the results of the second iteration for the top 2 and feature set combined with the top 2 feature. In this iteration, the most accurate feature sets, in order from most to least accurate, were Top 2 + centrality (accuracy = 0.885), Top 2+ idea embedding (accuracy = 0.547), by Top feature group (accuracy = 0.414), and Top 2 + sentiment (accuracy = 0.306).

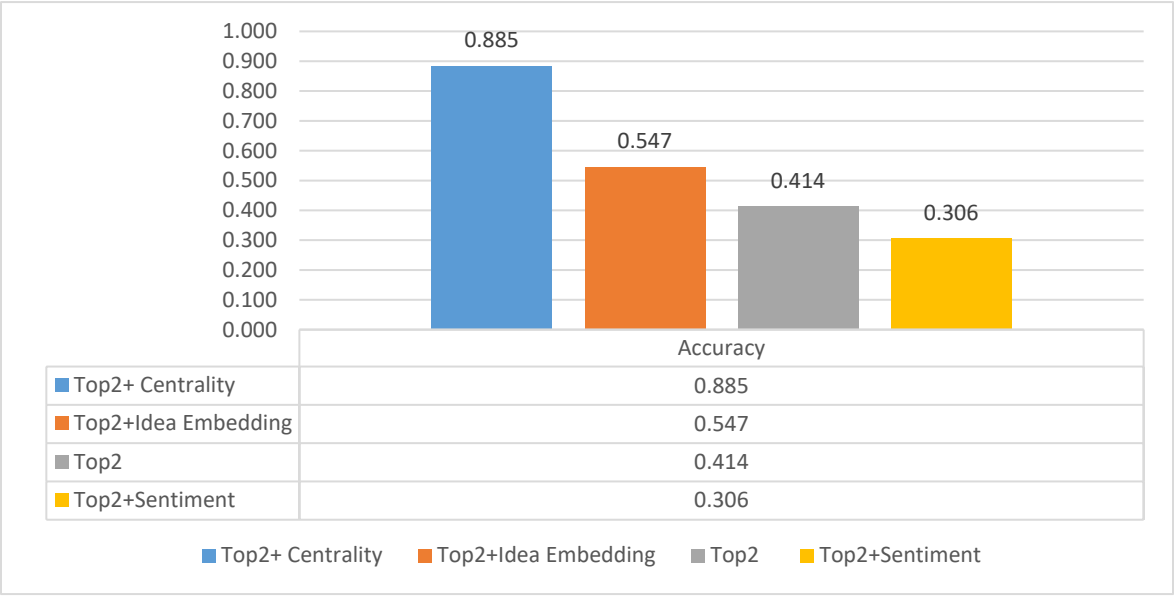

Figure 4.5 The Accuracy of Top2 and Combined Feature of SVM for Idea Popularity

For the third iteration, we combined the Top 2 + centrality as a new feature set called Top 3. The new feature sets also built from the Top 3 feature: Top 3 + idea embedding, Top 3 + sentiment, and Top 3. We then re-tested these three new feature sets using the SVM classifier to find the most accurate performance

Figure 4.6 summarizes the results of the fourth iteration for the top 3 and the feature set combined with top 3 feature group. In this iteration, the accuracy results are very close. The most accurate feature sets, in order from most to least accurate, were Top 3 + sentiment (accuracy = 0.890), Top 3 (accuracy = 0.885), Top 3 + idea embedding (accuracy = 0.878).

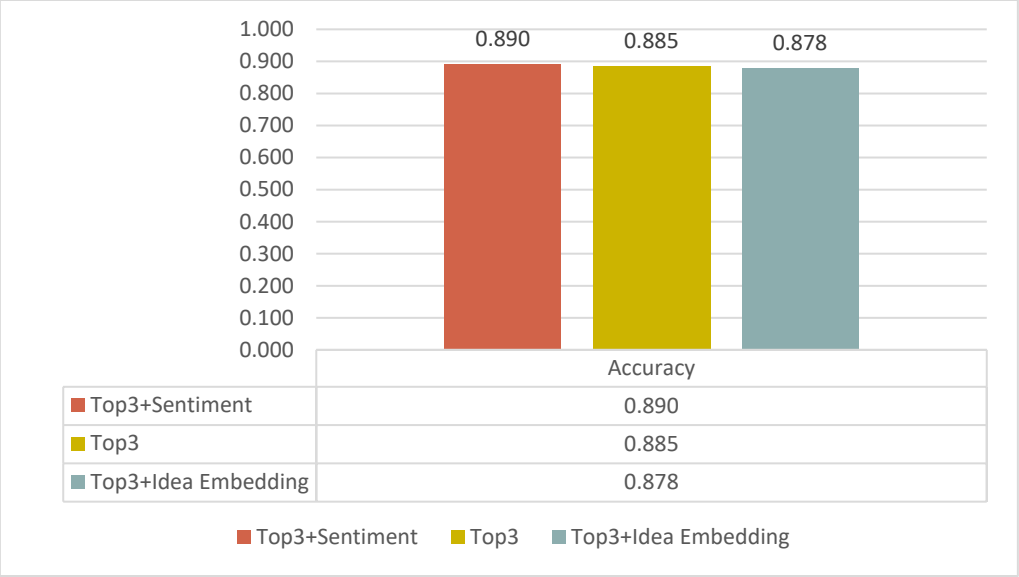

Figure 4.6 The Accuracy of Top3 and Combined Feature of SVM for idea popularity

In the final iteration, we combined Top 4 feature as a new feature set called Top 4 feature, containing the Top 3 and sentiment features. We used the SVM classifier to find the performance of all the features combined. The accuracy of all features for the SVM classifier is 0.890. In the final iteration, we combined all five features into a new feature set called All Feature, and we used the SVM classifier to find the performance of All Feature. The accuracy of All Feature for the SVM classifier is 0.786, which is lower than Top 4 in the last iteration.

Again, the findings indicate that accuracy was increased when we combined more features in the model. The Top 4 features that contributed the most to identifying the popularity of ideas were behavior, centrality, sentiment, and surface-level features. These results are in accord with previous studies that have used behavioral and textual features to predict classification. This research adds to previous studies by providing more detailed centrality measures and idea embedding approach.

4.4. The Classification of Individual Popularity by Idea Embedding, Surface-Level Features, Sentiment Features, Behavior Features, and Centrality Features

In this section, we detect users' popularity through user's Indegree measure. User popularity was treated as a binary classification (high and low, divided by the mean of all the users' Indegree). To detect individual popularity, each user's idea is characterized regarding surface-level feature group, sentiment feature group, and the behavior feature group listed in chapter 3, as well as the idea embedding. The goal is to compare their individual and combined performance to user popularity. Individuals' Indegree centrality is associated with centrality feature group, so centrality feature group was removed from this experiment, only used other three feature groups for the experiment and compared their individual and combined performances.

SVM was used in this experiment with cross-validation for the evaluation. The classification will be considered successful if the results are better than chance. Accuracy is used as the classification evaluation metric (Tang et al., 2009).

4.4.1 Individual Feature Selection

In the individual model, we separated all the single features for the individual popularity regression, to find the single feature importance by utilizing Random Forest Regression. The formula of Random Forest Regression is:

$$Regression = Bias + feature\_1\ contribution + \cdots + feature\_n\ contribution$$

This research adopted mean decrease impurity method, of Random Forest regression, where the optimal condition is chosen based on information gain for the regression model for the feature selection (Archer & Kimes, 2008; Guyon & Elisseeff, 2003; Menze et al., 2009; Peng et al., 2005; Saeys et al., 2008).

Table 4.8 The summary of feature importance for individual popularity by Random Forest Regression

Feature	**Contribution**
Number of submitted Ideas	0.643
Comment on own ideas	0.050
Idea Embedding 12th dimension	0.021
Idea length	0.014
Idea Embedding 27th dimension	0.013
Idea Embedding 17th dimension	0.010
Idea Embedding 99th dimension	0.009
Idea Embedding 37th dimension	0.008
Idea Embedding 93th dimension	0.008
Idea Embedding 16th dimension	0.007

According to the Table 4.8, The positive contribution to the individual popularity came from number of submitted ideas (removing it decreases model performance by 64.3%); Comment

on own ideas (removing it decreases model performance by 5%); one component of the idea embedding, corresponding to the 12th dimension of the idea embedding parameters; idea length (removing it decreases model performance by 1.4 %); and followed by six idea embedding components, corresponding to the 27th, 17th, 99th, 37th, 93th, 16th dimension of the idea embedding parameters (removing it decreases model performance by from 1.3 to 0.7 %). groups to compare the performance of the individual feature group instead of the entire feature set.

4.4.2 Individual Feature Group

Figure 4.7 summarizes the results of the first iteration. We tested each feature group separately over the ideas to examine which feature group is most important in identifying the popular individuals. We used the accuracy of SVM to examine the performance of other feature

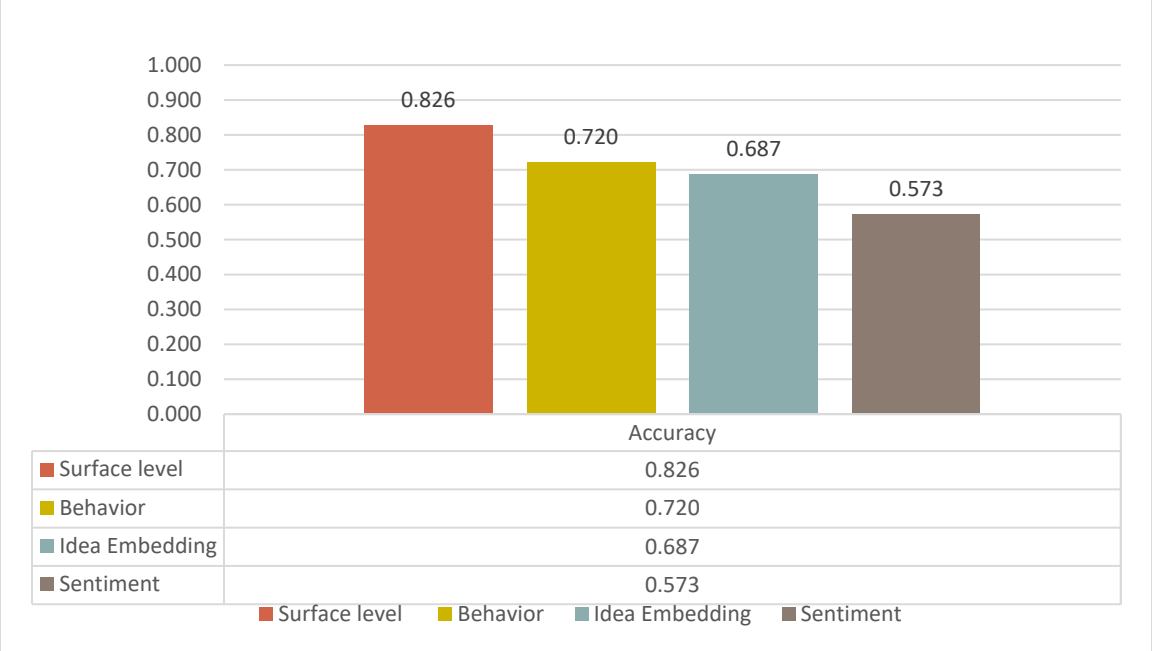

Figure 4.7 The Accuracy of Single Feature Group of SVM for Individual Popularity

As seen in Figure 4.7, the top 2 feature groups for identifying the popular individuals are surface-level of idea texts feature group (accuracy = 0.826) and behavior feature group (accuracy

= 0.720), followed by idea embedding (accuracy = 0.687), and sentiment feature group (accuracy = 0.573).

4.4.3 Feature Groups in Combinations

After we understood how each feature group performed individually, we also wanted to know what combination of feature groups would lead to the best performance for individual popularity detection. From the last section, we know that the Top 2 features of individual popularity classification are surface-level and behavior feature groups. They are most accurate in classifying individual popularity. So the next steps were to combine them into the same feature set, called the Top 2 feature. The other two new feature sets also built from the Top 2 feature: the Top 2 + sentiment feature group and Top 2 + idea embedding feature group.

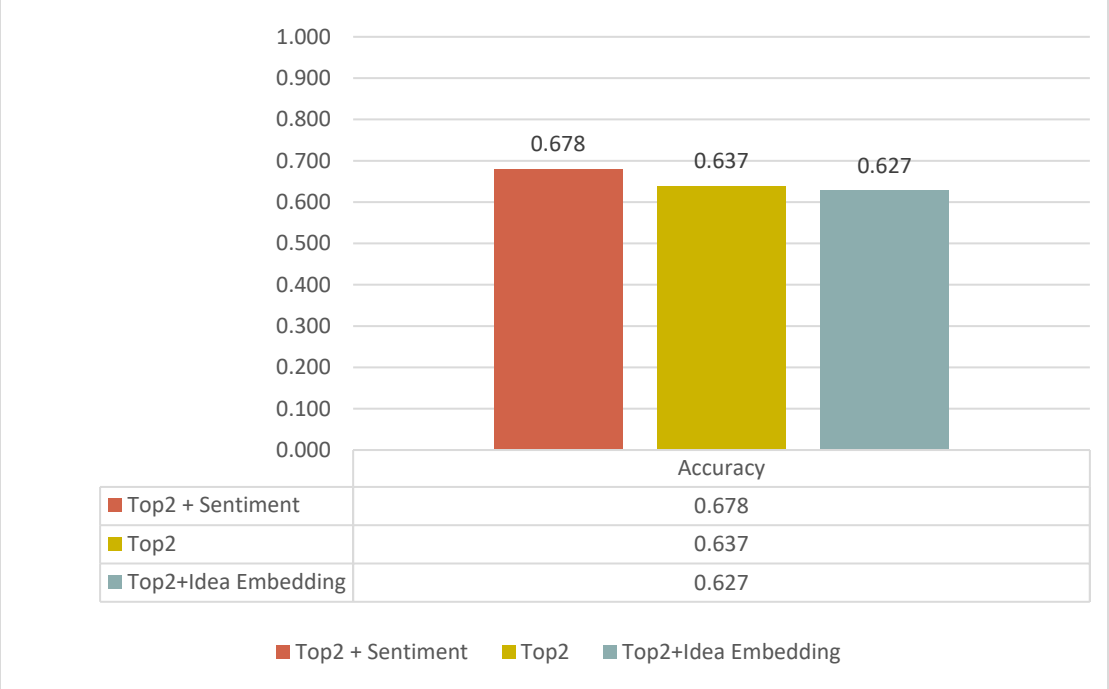

Figure 4.8 The Accuracy of Top2 and Combined Feature of SVM for Individual Popularity

The results of the second iteration for the Top 2 and the feature set combined with Top 2 features are as follows. The most accurate feature sets in order from the most to the least accurate were Top 2 + sentiment (accuracy = 0.678), Top 2 by itself (accuracy = 0.637), and finally Top 2 + idea embedding (accuracy = 0.627).

For the third iteration, we combined the Top 2 + sentiment as a new feature set called Top 3; the accuracy of this set was 0.678. In the final iteration, we combined all four features into a new feature set called Top 4 feature, which contains the Top 3 feature and idea embedding feature. The accuracy of all features combined for the SVM classifier was 0.645, which was lower than Top 3 (accuracy = 0.678) in the last iteration.

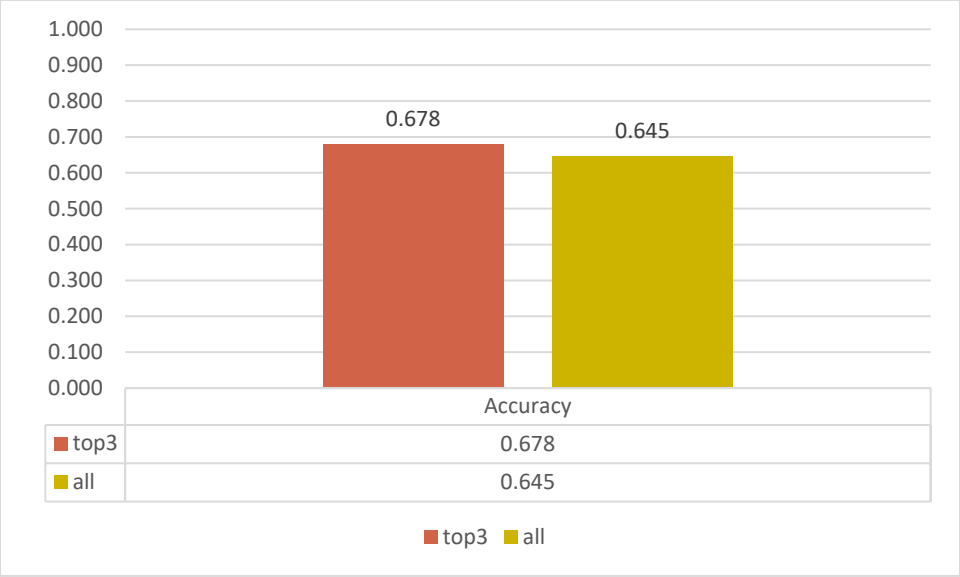

Figure 4.9 The Accuracy of Top3 and all Combined Feature of SVM for Individual Popularity

The findings for individual popularity show that accuracy was not always increased when we combined more features in the model. Also, the Top 3 features that contributed the most to identifying the popular individual were surface-level feature groups of idea texts, behavior feature

group, and sentiment. This section has attempted to use textual and behavioral features to predict individual popularity which was calculated by users' indegree. The results presented here are in accord with the results of previous studies that have used behavioral and textual features to predict user information. This finding enhances previous studies by providing more detailed textual feature groups for individual popularity detection.

CHAPTER 5. CONCLUSIONS

In this dissertation, we presented the results of detecting the adoptable ideas, the popularity of ideas, and the popularity of individuals by using the idea embedding, the surface-level features of online texts, the sentiment features of idea texts, the features of online behaviors, and the centrality features in the *MyStarbucksIdea* online community. This section reviews the empirical findings and then discusses the significance of the findings, research limitations, and suggestions for future research.

Research question 1 asked how acurately can idea adoption be classified based only on the idea embedding of idea texts. Because the percentage of the adopted ideas is only 4.6% of all the ideas, the classification is very skewed. Thus, we first used four resampling methods for imbalanced class adjustment, including random over-sampling, random under-sampling, SMOTe + ENN, and SMOTe + Tomek Links, to determine which resampling method would produce the best result for classification of the *MyStarbucksIdea* dataset. Additionally, three supervised machine learning approaches were employed as the classifiers, including Logistic Regression, SVM, and Random Forest, to compare the performances of different combinations of adjustment methods and classifiers.

After the adjustments due to the imbalanced classes, it was determined that the SMOTe + ENN resampling method produced the best results for Logistic Regression and SVM, followed by SMOTe + Tomek Links and Random Over-sampling. However, for the Random Forest classifier, Random Over-sampling is the most robust resampling method for producing the best results. These findings may be explained by considering SMOTe + ENN, which not only replicated the positive samples but also served as the data cleaning method to remove the samples between two classes and produced better results for the classifiers. However, because Random Forest is an ensemble

method, it can gather all the “weak learners” and help them become “strong leaners.” So Random over-sampling randomly replicated the positive samples, providing the Random Forest classifier with more information for correctly classifying the positive sample.

Besides the results of comparing the different resampling methods for imbalanced class adjustments, the results of different classifiers showed that the Random Forest had higher performance in evaluation metrics than Logistic Regression and SVM. The result showed, for idea adoption classification by the idea embedding, after the imbalanced class adjustment, the performance of Random Forest Classifier was over 95% in accuracy, AUC, and F-measure; the performance of Logistic regression was around 70 % in accuracy, AUC, and F-measure; and the performance of SVM was about 60% in accuracy, AUC, and F-measure.

The classification of idea adoption by using the idea embedding of online texts was considered successful because no matter which combination of resampling methods and classifiers were used, the chances were always better than 50%. The results showed that the idea embedding, which represents the overall meaning of ideas, was an important feature to include when classifying adopted and non-adopted ideas. This result complements previous studies that used features of online texts to predict user information or online review rating by contributing the new empirical findings of the importance of using the overall meaning of the online texts as a feature for classification.

Besides using the idea embedding as a feature, in research question 2, we explored how accurately idea adoption could be classified by the idea embedding, along with other features of online texts and online behaviors. After the imbalanced class adjustment, the top 10 single features for idea adoption classification were Outdegree centrality, Eigenvector centrality, number of submitted ideas, Indegree centrality, Degree centrality, Eccentricity centrality, Betweenness

centrality, and three components of the idea embedding vectors, corresponding to the 29th, 100th, and 54th dimension of the idea embedding parameters. In the top 10 single features, Outdegree centrality, Eigenvector centrality, Indegree centrality, Degree centrality, Eccentricity centrality, Betweenness centrality, were considered as centrality feature group, the frequency users submitted ideas was behavioral feature, and the others are idea embedding parameters. The result showed that idea embedding features and centrality features were most important for detecting adopted ideas.

Besides single features, we also compared the performance of classification of idea adoption by using feature groups. The results of surface-level feature group, Random Forest was found to have the highest performance on all the evaluation metrics: over 0.99 for accuracy, AUC, and F-measure. Logistic Regression performed at over 80% for accuracy, AUC, and F-measure. SVM performed at over 70% for accuracy, AUC, and F-measure. The classification by surface-level feature group was considered successful because all the results are better than chance. For the sentiment feature group, Random Forest was still found to have the highest performance on all the evaluation metrics. After the imbalanced class adjustment, the performance was over for 93% accuracy, AUC, and F-measure for Random Forest Classifier. For Logistic regression, the performance was over 50% for accuracy, AUC, and F-measure. The performance of SVM was over 50% for accuracy, AUC, and F-measure. The classification by sentiment feature group was considered successful because all the results are better than chance. For the behavioral feature group, Random Forest was still found to have the highest performance on all the evaluation metrics. After the imbalanced class adjustment, the performance was over for 98% accuracy, AUC, and F-measure for Random Forest Classifier. For Logistic regression, the performance was over 70% for accuracy, AUC, and F-measure. The performance of SVM was over 70% for accuracy, AUC, and

F-measure. The classification by behavior feature group was considered successful because all the results are better than chance. For the centrality feature group, Random Forest was still found to have the highest performance on all the evaluation metrics. After the imbalanced class adjustment, the performance was over for 99% accuracy, AUC, and F-measure for Random Forest Classifier. For Logistic regression, the performance was over 70% for accuracy, AUC, and F-measure. The performance of SVM was over 70% for accuracy, AUC, and F-measure. The classification by centrality feature group was considered successful because all the results are better than chance.

The findings of feature combinations for the classification of idea adoption indicate that accuracy was increased when we combined more feature groups in the model. These results are in accord with the results of the previous studies, which used behavioral and textual features to predict the classification and enhance the previous studies' results by providing the idea embedding approach and a more detailed centrality feature group.

Research question 3 asked how accurately the popularity of ideas could be classified by the idea embedding and the features of online texts and online behaviors. After the imbalanced class adjustment, the top 10 single features for the regression of idea popularity were Eigenvector centrality; comment on own ideas; one of the idea embedding components, corresponding to the 49^{th} dimension of the idea embedding parameters ; clustering coefficient; another one of the idea embedding components, corresponding to the 15^{th} dimension of the idea embedding parameters; the number of submitted ideas; and four components of the idea embedding, corresponding to the 88^{th}, 70^{th}, 9^{th}, 69^{th} dimension of the idea embedding parameters. The results showed that idea embedding features and centrality features were the most important features for detecting idea popularity.

For feature groups, the results showed the important feature group of identifying the popular ideas as behavior feature group (accuracy = 0.71), surface-level feature group (accuracy = 0.706), followed by sentiment feature group (accuracy = 0.704), idea embedding feature group (accuracy = 0.543), and centrality feature group (accuracy = 0.525). We know how each feature group performed individually; we also wanted to know what combination of features would lead to the best performance for idea popularity prediction. The findings indicated that accuracy was increased when we combined more feature groups in the model. The Top 4 feature groups that contributed the most to the performance of identifying popular ideas were behavior, centrality, sentiment, and surface-level feature groups. These results were in accord with the results of previous studies, which used the behavioral and textual features to predict the ratings of online reviews and enhance previous studies' results by providing more detailed centrality measures and latest document embedding approach.

Research question 4 considered how accurately the popularity of individual could be classified by the idea embedding of idea texts, surface-level features of idea texts and online behaviors. After the imbalanced class adjustment, the top 10 single features for individual popularity identification were the frequency of submitted ideas (removing it decreases model performance by 64.6%); comment on own ideas; idea length; and seven components of idea embedding, corresponding to the 12^{th}, 27^{th}, 17^{th}, 99^{th}, 37^{th}, 93^{th}, 16^{th} dimension of the idea embedding parameters. The results showed that the idea embedding features and behavioral features are most important for detecting the popularity of individuals.

For feature combinations, the Top 2 feature groups for the individual popularity detection were surface-level feature group and behavior feature group. They were the most accurate in classifying individual popularity. The findings for the classification of popularity of individuals

showed that accuracy was not always increased when we combined more features in the model. Moreover, the top three feature group that contributed the most to identifying the popularity of individuals were surface-level of idea texts, behavior, and sentiment feature group.

It is worth mentioning that the conclusions drawn above should be interpreted in relation to the specific context of *MyStarbucksIdea* online platform. These findings are in line with previous studies, although previous research has used different methods to extract the features in online texts and behaviors.

We acknowledge that this research is exploratory and that there are some methodological limitations in the research design that limit the interpretations. Even though the research has the merit of offering valuable insights into the combination of Doc2Vec with other features, we used 100-dimensional vectors for idea representation. So if we combine Doc2Vec with other feature groups and feed them into the regression model, Doc2Vec is not treated as a single feature; instead, each of the 100 dimensions counts as one feature. Thus, in the regression model, the performance of idea embedding, or the overall meaning of the ideas, could not represent the overall performance of idea embedding, but rather that of a single component of a vector. Additionally, the method we used to combine the feature groups was transforming a feature group as a matrix, and then combined each matrix. If there is a correlation between two matrices, the performance of combined feature groups may be affected by that. However, we could not well explain this mechanism at this moment.

The second limitation concerns imbalanced classes. Regardless of whether or not ideas were adopted, idea popularity and individual popularity data are all imbalanced classifications within the small positive samples. So all ideas in the *MyStarbucksIdea* dataset were used as the trained data for classification and detection. There is no new data for prediction because there are

not enough positive samples to be split into prediction datasets. Therefore, the prediction could not be implemented.

The third limitation concerns the quality of the dataset. The sparsity of online texts affected the number of meaningful connections among the words, which may have influenced the performance of the representation of ideas by the Doc2Vec, especially since we used the PV-DM model to preserve word order in the ideas.

The fourth limitation concerns centrality measures based on online behavior. The centrality measure in this research was based on commenting behavior. However, while some individuals' comment on others' ideas at different time points and over an interval of years, others comment on ideas within a very short period. This disparity between infrequent and constant interaction may affect impact significance, but we could not measure this in the present study.

Thus, based on the limitations of this study, the generalization of the results to other online forums may be limited.

For social media or any online forum, we can mainly capture textual data and behavioral data. The methodological question raised here is that how we can combine both textual and behavioral data appropriately for analysis. This research takes the first step in answering this question, and it is hoped that future work will clarify these reliability and validity concerns. It remains for future research to account for solving the concerns of feature combinations for machine learning algorithms. This research presented the preliminary results of pilot experiments that will need to be further analyzed, expanded and replicated, and more work in this area will make the detection and prediction more precise. It is hoped that the findings of this research will serve as a basis for further study.

REFERENCES

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013). *Empirical evaluation of profile characteristics for gender classification on twitter*. Paper presented at the Machine Learning and Applications (ICMLA), 2013 12th International Conference on.

Amrit, C., & Van Hillegersberg, J. (2010). Exploring the impact of socio-technical core-periphery structures in open source software development. *journal of information technology, 25*(2), 216-229.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119-123.

Asoh, H., Ikeda, K., & Ono, C. (2012). *A Fast and Simple Method for Profiling a Population of Twitter Users.* Paper presented at the The Third International Workshop on Mining Ubiquitous and Social Environments.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135-160.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter, 6*(1), 20-29.

Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management science, 59*(1), 226-244.

Benton, A., Mitchell, M., & Hovy, D. (2017). *Multi-task learning for mental health using social media text.* Paper presented at the Proceedings of EACL.

Beretta, V., Maccagnola, D., Cribbin, T., & Messina, E. (2015). *An interactive method for inferring demographic attributes in Twitter*. Paper presented at the Proceedings of the 26th ACM Conference on Hypertext & Social Media.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: " O'Reilly Media, Inc.".

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Grobler, J. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Cesare, N., Grant, C., & Nsoesie, E. O. (2017). Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices. *arXiv preprint arXiv:1702.01807*.

Chan, K. W., & Li, S. Y. (2010). Understanding consumer-to-consumer interactions in virtual communities: The salience of reciprocity. *Journal of Business Research, 63*(9-10), 1033-1040.

Chen, L., Qi, L., & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications, 39*(10), 9588-9601.

Cheung, C. M., & Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision support systems, 53*(1), 218-225.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Icwsm, 133*, 89-96.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). *From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses.* Paper presented at the CLPsych@ HLT-NAACL.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). *CLPsych 2015 Shared Task: Depression and PTSD on Twitter.* Paper presented at the CLPsych@ HLT-NAACL.

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). *Exploratory Analysis of Social Media Prior to a Suicide Attempt.* Paper presented at the CLPsych@ HLT-NAACL.

Cross, R., Laseter, T., Parker, A., & Velasquez, G. (2006). Using social network analysis to improve communities of practice. *California Management Review, 49*(1), 32-60.

Crowston, K., Wei, K., Li, Q., & Howison, J. (2006). *Core and periphery in free/libre and open source software team communications.* Paper presented at the System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on.

Culotta, A., Kumar, N. R., & Cutler, J. (2015). *Predicting the Demographics of Twitter Users from Website Traffic Data.* Paper presented at the AAAI.

David, E., David, E., Zhitomirsky-Geffet, M., Zhitomirsky-Geffet, M., Koppel, M., Koppel, M., . . . Uzan, H. (2016). Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. *Online Information Review, 40*(5), 610-623.

Duan, W., Cao, Q., & Gan, Q. (2010). *Investigating Determinants of Voting for the" Helpfulness" of Online Consumer Reviews: A Text Mining Approach.* Paper presented at the AMCIS.

Dugosh, K. L., & Paulus, P. B. (2005). Cognitive and social comparison processes in brainstorming. *Journal of experimental social psychology, 41*(3), 313-320.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science, 38*(2), 189-200. doi:10.1177/0165551512437638

Faraj, S., & Johnson, S. L. (2011). Network exchange patterns in online communities. *Organization Science, 22*(6), 1464-1480.

Feng, H., & Lin, R. (2016). Sentiment Classification of Food Reviews. *arXiv preprint arXiv:1609.01933*.

Filippova, K. (2012). *User demographics and language in an implicit social network.* Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Fink, C., Kopecky, J., & Morawski, M. (2012). *Inferring Gender from the Content of Tweets: A Region Specific Example.* Paper presented at the Icwsm.

Fonti, F., & Maoret, M. (2016). The direct and indirect effects of core and peripheral social capital on organizational performance. *Strategic Management Journal, 37*(8), 1765-1786.

Fuger, S., Schimpf, R., Füller, J., & Hutter, K. (2017). User roles and team structures in a crowdsourcing community for international development–a social network perspective. *Information Technology for Development*, 1-25.

Füller, J., Hutter, K., Hautz, J., & Matzler, K. (2014). User roles and contributions in innovation-contest communities. *Journal of Management Information Systems, 31*(1), 273-308.

Füller, J., Jawecki, G., & Mühlbacher, H. (2007). Innovation creation by online basketball communities. *Journal of Business Research, 60*(1), 60-71.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498-1512.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). *Predicting personality from twitter.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.

Golbeck, J., Robles, C., & Turner, K. (2011). *Predicting personality with social media.* Paper presented at the CHI'11 extended abstracts on human factors in computing systems.

Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J.-P., Sanchez-Perez, M. A., & Chanona-Hernandez, L. (2016). Improving feature representation based on a neural network for author profiling in social media texts. *Computational intelligence and neuroscience, 2016*, 2.

Goswami, S., Sarkar, S., & Rustagi, M. (2009). *Stylometric analysis of bloggers' age and gender.* Paper presented at the Third International AAAI Conference on Weblogs and Social Media.

Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Web semantics: science, services and agents on the World Wide Web, 6*(1), 4-13.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Hopcroft, J. E., & Tarjan, R. E. (1973). Dividing a graph into triconnected components. *SIAM journal on computing, 2*(3), 135-158.

Hossain, M., & Islam, K. Z. (2015). Generating ideas on online platforms: A case study of "My Starbucks Idea". *Arab Economic and Business Journal, 10*(2), 102-111.

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior, 48*, 17-27.

Ikeda, D., Takamura, H., & Okumura, M. (2008). *Semi-Supervised Learning for Blog Classification.* Paper presented at the AAAI.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science, 30*(2), 195-212.

Jalili, M., Salehzadeh-Yazdi, A., Asgari, Y., Arab, S. S., Yaghmaie, M., Ghavamzadeh, A., & Alimoghaddam, K. (2015). CentiServer: a comprehensive resource, web-based application and R package for centrality analysis. *PloS one, 10*(11), e0143111.

Jalili, M., Salehzadeh-Yazdi, A., Gupta, S., Wolkenhauer, O., Yaghmaie, M., Resendis-Antonio, O., & Alimoghaddam, K. (2016). Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Frontiers in physiology, 7*, 375.

Jin, Z., Li, Q., Zeng, D. D., Zhan, Y., Liu, R., Wang, L., & Ma, H. (2016). *Jointly modeling review content and aspect ratings for review rating prediction.* Paper presented at the Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65): Springer.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802-5805.

Kuppuswamy, V., & Bayus, B. L. (2015). Crowdfunding creative ideas: The dynamics of project backers in Kickstarter.

Lam, S. S., & Schaubroeck, J. (2000). A field experiment testing frontline opinion leaders as change agents. *Journal of Applied Psychology*, *85*(6), 987.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents.* Paper presented at the International Conference on Machine Learning.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, *18*(17), 1-5.

Lévy, P. (1997). *Collective intelligence*: Plenum/Harper Collins New York.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, *109*(1), 68-72.

Li, F., & Du, T. C. (2011). Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. *Decision support systems*, *51*(1), 190-197.

Liu, W., & Ruths, D. (2013). *What's in a Name? Using First Names as Features for Gender Inference in Twitter*. Paper presented at the AAAI spring symposium: Analyzing microtext.

Ludu, P. S. (2014). Inferring gender of a Twitter user using celebrities it follows. *arXiv preprint arXiv:1405.6667*.

Malhotra, A., & Majchrzak, A. (2014). Managing crowds in innovation challenges. *California Management Review*, *56*(4), 103-123.

Malouf, R., & Mullen, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research, 18*(2), 177-190.

Manchanda, P., Packard, G., & Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science, 34*(3), 367-387.

Markov, I., Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., & Gelbukh, A. (2016). *Author profiling with doc2vec neural network-based document embeddings.* Paper presented at the Mexican International Conference on Artificial Intelligence.

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.-F., Davalos, S., Teredesai, A., & De Cock, M. (2014). *Age and gender identification in social media.* Paper presented at the Proceedings of CLEF 2014 Evaluation Labs.

Mechti, S., Jaoua, M., Belguith, L. H., & Faiz, R. (2014). Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media. *Proceedings of the PAN@ CLEF, Sheffield, England*.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics, 10*(1), 213.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model.* Paper presented at the Interspeech.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). *Linguistic regularities in continuous space word representations.* Paper presented at the Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science, 2*(04), 143.

Mohammady, E., & Culotta, A. (2014). Using county demographics to infer attributes of twitter users. *ACL 2014, 7*.

Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly*, 185-200.

Ngo-Ye, T. L., & Sinha, A. P. (2012). Analyzing online review helpfulness using a regressional ReliefF-enhanced text mining method. *ACM Transactions on Management Information Systems (TMIS), 3*(2), 10.

Nguyen, D.-P., Gravel, R., Trieschnigg, R. B., & Meder, T. (2013). " How old do you think I am?" A study of language and age in Twitter.

Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T., & Yeung, C.-M. A. (2013). TweetGenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter, 4*(4).

Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). *Author age prediction from text using linear regression.* Paper presented at the Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.

Nguyen, T., Phung, D. Q., Adams, B., & Venkatesh, S. (2011). *Prediction of Age, Sentiment, and Connectivity from Social Media Text.* Paper presented at the WISE.

Nisbet, M. C., & Kotcher, J. E. (2009). A two-step flow of influence? Opinion-leader campaigns on climate change. *Science Communication, 30*(3), 328-354.

Niu, L., Dai, X., Zhang, J., & Chen, J. (2015). *Topic2Vec: learning distributed representations of topics.* Paper presented at the Asian Language Processing (IALP), 2015 International Conference on.

Nolker, R. D., & Zhou, L. (2005). *Social computing and weighting to identify member roles in online communities.* Paper presented at the Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on.

Nowson, S., & Oberlander, J. (2006). *The Identity of Bloggers: Openness and Gender in Personal Weblogs.* Paper presented at the AAAI spring symposium: Computational approaches to analyzing weblogs.

Park, C. S. (2013). Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement. *Computers in Human Behavior, 29*(4), 1641-1648.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825-2830.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). *Predicting age and gender in online social networks.* Paper presented at the Proceedings of the 3rd international workshop on Search and mining user-generated contents.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27*(8), 1226-1238.

Pennacchiotti, M., & Popescu, A.-M. (2011). *Democrats, republicans and starbucks afficionados: user classification in twitter.* Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., . . . Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights, 5*(Suppl 1), 3.

Plank, B., & Hovy, D. (2015). *Personality Traits on Twitter-or-How to Get 1, 500 Personality Tests in a Week.* Paper presented at the WASSA@ EMNLP.

Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015). *An analysis of the user occupational class through Twitter content*.

Ramaswamy, V., & Gouillart, F. (2010). Building the co-creative enterprise. *Harvard business review, 88*(10), 100-109.

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. *Icwsm, 11*, 598-601.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying latent user attributes in twitter.* Paper presented at the Proceedings of the 2nd international workshop on Search and mining user-generated contents.

Reddy, S., Wellesley, M., Knight, K., & del Rey, C. M. (2016). Obfuscating gender in social media writing. *NLP+ CSS 2016*, 17.

Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora.* Paper presented at the In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Rochat, Y. (2009). *Closeness centrality extended to unconnected graphs: The harmonic centrality index.* Paper presented at the ASNA.

Rosenthal, S. (2014). Detecting influencers in social media discussions. *XRDS: Crossroads, The ACM Magazine for Students, 21*(1), 40-45.

Rosenthal, S., & McKeown, K. (2011). *Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Rosenthal, S., & McKeown, K. (2016). Social proof: The impact of author traits on influence detection. *NLP+ CSS 2016*, 27.

Rowe, M., & Strohmaier, M. (2014). *The semantic evolution of online communities.* Paper presented at the Proceedings of the 23rd International Conference on World Wide Web.

Rullani, F., & Haefliger, S. (2013). The periphery on stage: The intra-organizational dynamics in online communities of creation. *Research Policy, 42*(4), 941-953.

Rustagi, M., Prasath, R. R., Goswami, S., & Sarkar, S. (2009). Learning Age and Gender of Blogger from Stylistic Variation. *PReMI, 9*, 205-212.

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). *Robust feature selection using ensemble feature selection techniques.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Santosh, K., Joshi, A., Gupta, M., & Varma, V. (2014). *Exploiting Wikipedia Categorization for Predicting Age and Gender of Blog Authors.* Paper presented at the UMAP Workshops.

Setia, P., Rajagopalan, B., Sambamurthy, V., & Calantone, R. (2012). How peripheral developers contribute to open-source software development. *Information Systems Research, 23*(1), 144-163.

Siswanto, E., & Khodra, M. L. (2013). *Predicting latent attributes of Twitter user by employing lexical features.* Paper presented at the Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on.

Stuart, L. M., Tazhibayeva, S., Wagoner, A. R., & Taylor, J. M. (2013). *On identifying authors with style.* Paper presented at the Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on.

Surowiecki, J. (2005). *The wisdom of crowds*: Anchor.

Tang, Y., Zhang, Y.-Q., Chawla, N. V., & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 281-288.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing, 1*(2), 146-160.

Tomek, I. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics, 6*, 769-772.

Toral, S. L., Martínez-Torres, M. d. R., & Barrero, F. (2010). Analysis of virtual communities supporting OSS projects using social network analysis. *Information and Software Technology, 52*(3), 296-303.

Tuli, G. (2016). Modeling and Twitter-based Surveillance of Smoking Contagion.

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). *Inferring Latent User Properties from Texts Published in Social Media.* Paper presented at the AAAI.

Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, 35-57.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature, 393*(6684), 440.

Yao, R., & Chen, J. (2013). *Predicting movie sales revenue using online reviews.* Paper presented at the Granular Computing (GrC), 2013 IEEE International Conference on.

Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720-734.

Zhang, C., & Zhang, P. (2010). Predicting gender from blog posts. *University of Massachussetts Amherst, USA.*