RETWEET PREDICTION BASED ON USER BEHAVIOR

By

Syeda Nadia Firdaus

Master of Science in Computer Science

Ryerson University, Toronto, Canada, 2013

Bachelor of Science in Computer Science and Engineering

The University of Asia Pacific, Bangladesh, 2005

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Computer Science

Toronto, Canada, 2019

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

RETWEET PREDICTION BASED ON USER BEHAVIOR

By

Syeda Nadia Firdaus

Doctor of Philosophy in Computer Science

Ryerson University, Toronto, Canada, 2019

Abstract

Social network is a hot topic of interest for researchers in the field of computer science in recent years. These social networks such as Facebook, Twitter, Instagram play an important role in information diffusion. Social network data are created by its users. Users' online activities and behavior have been studied in various past research efforts in order to get a better understanding on how information is diffused on social networks. In this study, we focus on Twitter and we explore the impact of user behavior on their retweet activity. To represent a user's behavior for predicting their retweet decision, we introduce 10-dimentional emotion and 35-dimensional personality related features. We consider the difference of a user being an author and a retweeter in terms of their behaviors, and propose a machine learning based retweet prediction model considering this difference. We also propose two approaches for matrix factorization retweet prediction model which learns the latent relation between users and tweets to predict the user's retweet decision. In the experiment, we have tested our proposed models. We find that models based on user behavior related features provide good improvement (3% - 6% in terms of F1-score) over baseline models. By only considering user's behavior as a retweeter, the data processing time is reduced while the prediction accuracy is comparable to the case when both retweeting and posting behaviors are considered. In the proposed matrix factorization models, we

include tweet features into the basic factorization model through newly defined regularization terms and improve the performance by 3% - 4% in terms of F1-score. Finally, we compare the performance of machine learning and matrix factorization models for retweet prediction and find that none of the models is superior to the other in all occasions. Therefore, different models should be used depending on how prediction results will be used. Machine learning model is preferable when a model's performance quality is important such as for tweet re-ranking and tweet recommendation. Matrix factorization is a preferred option when model's positive retweet prediction capability is more important such as for marketing campaign and finding potential retweeters.

Acknowledgements

I would like to express sincere gratitude to my supervisors Dr. Cherie Ding and Dr. Alireza Sadeghian for their continuous support, encouragement, and time throughout the five years of doctoral studies. They have guided me with patience and care to improve my work in thesis. Without their kind cooperation and support, completion of research and publication in journals and conference would not be possible. Working under the supervision of Dr. Ding and Dr. Sadeghian has been a great and memorable experience for me.

I would like to thank my thesis committee members Dr. Eric Harley, Dr. Ebrahim Bagheri, and Dr. Marek Reformat for their time, patience and proficiency in judging my thesis. Their valuable comments were very helpful to improve my work.

I would also like to convey my gratitude to the faculty members of the Department of Computer Science, Ryerson University. Attending courses under guidance of committed professors helped me to advance my knowledge in computer science. My sincere appreciation goes to staff members of Computer Science department and fellow graduate students for their continuous support over the last five years.

I would like to thank Ryerson University Graduate Study to award me Ryerson graduate scholarship which helped me to pay my tuition fees for the graduate program, and NSERC for research stipends I received during my doctoral study period, which were a great financial help for me.

Lastly, I am extremely grateful for the support and inspiration of my family. Without them I would not be able to attain my goal and fulfill my dream to work in my field of interest. No specific word of appreciation would be enough to convey my love and thankfulness towards them.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

List of Acronyms

ALS	Alternating Least Square
ANEW	Affective Norms of English Words
API	Application Program Interface
APL	Average Path Length
AUC	Area Under Curve
AUPRC	Area Under Precision Recall Curve
CRF	Condition Random Field
HDP	Hierarchical Dirichlet Process
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
Mallet	Machine Learning for Language Toolkit
MAP	Mean Average Precision
PA	Passive-Aggressive
REST	Representational State Transfer
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
RWS	Random Walk with Restart
SGD	Stochastic Gradient Descent
SMO	Sequential Minimal Optimization
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
XGBoost	Extreme Gradient Boosting

List of Appendices

Chapter 1

Introduction

Human beings tend to be social in nature. Being social not only means to live in a society but also means to exchange views and information with members of society. Traditional social network represents a society in a specific geographical location, or serving a specific purpose (e.g., common interest). The Internet helps us form society including people from all over the world and serves all kinds of purposes. In simple words, online social networks are networks of people formed over the Internet based on social and professional relation, professional or personal interest, and social and humanitarian grounds. It carries a great amount of data which reflect its users' interest, behavior, and activities. Since this network does not rely on direct contact, sometimes it carries more in-depth information about people's opinions and activities, specifically on prominent and sensitive issues (e.g., people's opinion and preference for upcoming election, people's attitude and opinion towards terrorism). Online social network can be considered as a great repository of information. It also has the ability to spread information all around the world in the least amount of time. The data collected from these social networks have the potential to make an effective contribution in many different areas of research, such as marketing, business analysis, and human psychology analysis. One of the important areas of research that is using social network data is the study of social networks as mechanism for information diffusion. This research works on Twitter's information diffusion mechanism known as *Retweet*.

1.1 Background

In today's world, there are different social networking services. Each of these services provides unique facilities. Users have the opportunity to choose and use their preferred social networks. Every social networking service has their exclusive information spreading mechanism. We have decided to work with Retweet which is an important information diffusion mechanism provided by Twitter social networking service. We have selected Twitter because of its popularity. As an easily accessible service, Twitter gained its popularity all over the world. Twitter allows its users to create profile, publish messages, and share information with others. On Twitter, users' posts are known as tweets which are not more than 140 characters long. These tweets may contain URLs, hashtags (keywords followed by "#" symbol used to categorize the tweet), mentions (other users' usernames followed by "@" symbol), and emoticons. Users can also include photos and videos in their tweets. In Twitter, there exists follower–followee relationship between users. When a user wants to subscribe to other users' posts, they can follow them. For example, if user *A* follows user *B, A* is known as follower of *B* and *B* is known as followee of *A.* Twitter allows its users to maintain one-way relationship; which means that user *A* can follow any other user whereas it is not mandatory that other users have to follow *A*. In this type of relationship, it is very typical to follow celebrities and famous people to get continuous update from them. When user *A* follows *B, B*'s posts will appear in *A'*s Twitter main page. *A* can also like and repost *B*'s tweets. These reposts are known as **retweets**. Briefly, retweets are the re-post of author's message by another user to make it visible to that user's followers. Retweets look like original tweets with keyword "RT" and author's username (followed by "@" symbol) at the beginning of the text. Since tweets are restricted to be 140 characters long, authors of the tweets put emphasis on the formation of the tweets to make them useful and understandable. Retweeting can be considered as a fast information diffusion process because the user only needs to repost another user's message without taking own time to organize the message. Research is being done to find the tweets which have the potential to be retweeted by the users. Since users are the main actors for posting and spreading messages through online social network, research related to retweet prediction also includes finding out potential retweeters as well as finding out potential tweets for recommending to users. The importance of retweets for the purpose of

information diffusion has made it a significant topic of research in the field of social data analytics.

1.2 Motivation and Problem Statement

Retweet prediction is an imperative area of research which mainly deals with the prediction of potential retweets. Prediction of retweets is important for finding users' preference for spreading information. Since retweeting service is widely used for digital marketing and social movement, it is important to find the target user's preference in this spectrum. Since social network data is generated by users and they play important role in the process of information diffusion, user behavior and activities are investigated to understand this process. Different explicit or implicit content-based features as well as features related to social connections between authors and users have been explored by past researchers for the purpose of retweet prediction (Macskassy & Michelson, 2011; Uysal & Croft, 2011; Chen et al., 2012; Naveed et al., 2011; Kim & Yoo, 2012; Petrovic et al., 2011; Xu & Yang, 2012; Yang et al., 2010; Pfitzner et al., 2012; Jenders et al., 2013; Hoang & Lim, 2013). Content-based features represent the information contained in the tweets whereas social connection related features mainly represent the influence of the author in the user's retweet decision. Explicit features are directly measurable such as hashtags, URLs, emoticons, punctuation marks, and usernames (Peng et al., 2011; Suh et al., 2010; Xu & Yang, 2012; Naveed et al., 2011; Uysal & Croft, 2011; Kim & Yoo, 2012). Implicit features are not directly measurable. Tools or algorithms are needed to extract these features. Popular implicit features related to tweet include topics, terms with their TF–IDF scores, topic novelty, sentiment, and emotional divergence (Naveed et al., 2011; Uysal & Croft, 2011; Petrovic et al., 2011; Xu & Yang, 2012; Yang et al., 2010; Pfitzner et al., 2012; Jenders et al., 2013; Hoang et al., 2013; Chen et al., 2012). The goal of this research is to explore the effect of users' behavioral pattern on their retweeting activity. This research proposes that users' behavioral pattern can be represented by features related to their interests and attitudes, which in general are implicit content-based features. To explore a user's interest, topics of their posts are examined as features. To explore a user's attitude, their emotion, sentiment, and personality reflected by the posts are included as features. A few explicit content features such as

hashtags, usernames, and URLs are also included because they have been used widely in past research works and have been shown as effective features for retweet prediction. The main purpose of this research is to investigate the impact of users' behavior related features on their retweet decision.

User behavioral patterns are not easily deducible, but they play very important role in users' posting behavior (Xu et al., 2012). Specially in online communities, users do not need to communicate or interact directly with one another and they do not need to have any face to face contact with other members, so some users might be able to express their opinion more confidently in a carefree manner. Users' retweeting decisions are dependent on many factors such as posting time of the tweet, topic of the tweet, author of the tweet, and retweeter's intention and interest. Impact of tweet's structure and content on the user's retweeting decision has been investigated vastly (Petrovic et al., 2011; Chen et al., 2012; Uysal & Croft 2011; Naveed et al., 2011; Uysal & Croft, 2011; Kim & Yoo 2012; Xu & Yang 20120; Yang et al., 2010; Pfitzner et al., 2012; Jenders et al., 2013; Hoang et al., 2013; Chen et al., 2012). Researchers showed that structure and content of tweet has good impact on its retweetability. Structure of tweet refers to its construction and composition such as length of the tweet, and number of words in the tweet. Content of the tweet refers to information contained in the tweet. It was also observed that the hidden information such as user's personality, sentiment, and topic preference have influence on their retweet decision. These factors are concealed but put subtle influence on users' retweet activity. The previous researchers investigated a few implicit features such as topic, sentiment, and personality (for finding retweeters) separately in different research (Naveed et al., 2011; Pfitzner et al., 2012; Stieglitz & Dang-Xuan, 2012; Jenders et al., 2013; Kim & Yoo 2012, Xu & Yang, 2012, Peng et al., 2011). This research is intended to build machine learning based retweet prediction model to explore the combination effects of all these hidden factors on users' retweet decision.

In this study, implicit content-based features such as the user's topic interest, emotion, and personality are chosen to represent their interest and attitude. The assumption is that topic, emotion, and personality have subtle but strong influence on users' retweet decision. Every user's retweet preference in terms of topic is different. But topic preference is not the sole

decision maker for retweet activity. Users might prefer a topic but they might not retweet a message if they do not support the emotion reflected by the tweet. A user might like a topic but they might not retweet the tweet if they do not share the personality reflected by the tweet. A user might retweet a tweet if they prefer a certain emotion and share similar personality reflected by the tweet no matter what the topic is. These factors combined together collectively influence users' retweet decision. This research shows the effect and combinatory impact of different implicit features on predicting potential retweets for users.

An author's behavioral pattern may not exactly match with their behavioral pattern as a retweeter. In case of retweet decision, users are not the author of the message. They are the spreader of the message. Emotion and personality reflected by their retweets might not be same as those reflected by their own posts. Retweet is the tweet that reflects author's emotion and personality and is preferred by the user. According to the past research, users' retweeting behavior is dependent on the following three factors: author of the tweet, user who reads/sees and retweets the tweet, and the content of the tweet. In most of the cases, retweet prediction is based on the matching score between the user profile and the target tweet profile, the matching score between the user profile and the author's profile, or the matching score between the user's preferred content and content of the target tweet. These prediction models are developed under the following assumptions: a user will retweet tweets from like-natured and like-minded people, a user's behavior as an author and retweeter is same, a user will retweet the tweets which belong to their topic of interest. People sometimes retweet a tweet after adding opposing comments to it. This type of retweets is out of the scope of this research. This research works from the viewpoint of people retweeting tweets when they agree with it. For retweet prediction research, a user's profile is built using their activities on Twitter, which includes mainly their tweets and retweets. Although most of the time these collective activities represent a user's overall preference and behavior, sometimes there can be a fine line between a user's behavior and activities as an author and as a retweeter. For example, a painter is very much fond of arts and creative topics and always tweets or retweets anything related to their own profession or interest. However, it does not guarantee that they would not like scientific topics. Maybe they do not tweet anything regarding science but they may like to read scientific articles and want to spread messages covering scientific topics. And similarly, it can be said that a person may be introvert and always

tweets in a manner which reflects their introversion nature, but they may like to retweet something opposite to their nature. In another example regarding a user's emotional behavior, it can be said that a person might not feel comfortable to show a certain emotion, for example disgust or anger on a topic through their posts; but they can express their emotion indirectly to the world by retweeting messages reflecting this emotion. We consider that a user's behavior is different as an author and a retweeter and a user's past retweets carry more information than his past tweets in case of retweet prediction. The conventional strategy assumes that a user's tweets and retweets both provide important information about their future activity. Conventionally, a user's profile is developed using both their past tweets and retweets. This study would like to find out a better way of building user profile among three options – using retweets only or using tweets only or using both retweets and tweets.

The purpose of this research is to develop a model which can predict users' retweet decision accurately. This can be considered as a binary classification problem which classifies a target tweet either as positive (user retweets the tweets) or negative (user sees the tweet but does not retweet it). More specifically, objective of this research is to develop a retweet prediction model based on users' behavioural patterns. A user's behavior is represented by their interest and attitude related features.

1.3 Objectives

The goal of this research is to develop retweet prediction model based on users' behavioral pattern which can predict their retweet decision more accurately than the past methods. More detailed objectives of this research are described as follows:

- Explore the impact of users' behavior on their retweet decision. More concretely, we want to investigate the effect of the proposed implicit content-based features (topic, emotion, personality) along with previously used explicit content-based features (URL, hashtags, user-mention) to check whether the added implicit content-based features have any improvement over the previously used explicit content-based features. Also, compared to previous studies (Naveed et al., 2011; Pfitzner et al., 2012; Stieglitz &

Dang-Xuan, 2012; Jenders et al., 2013; Kim & Yoo, 2012), this research included a more complete list of emotion-sentiment features to check their impact on the accuracy of retweet prediction.

- Study the performance of different combinations of implicit content-based features to investigate the impact of these individual features related to users' interest and attitude on their retweet decision.

- Use machine learning technique to develop the retweet prediction models using the above-mentioned features because machine learning is the most common strategy to develop prediction models.

- Check whether we can achieve better or comparable results if we consider that a user's behavior is different as an author and retweeter. We want to examine a user's different behavior in different role and investigate the performance of retweet prediction models developed considering the following three hypotheses: a user's future retweets are similar to their past tweets and retweets; a user's future retweets are similar to their past retweets only; a user's future retweets are similar to their past tweets only.

- Explore the performance of matrix factorization technique to build retweet prediction model. Matrix factorization is a well-established and popular method to develop recommender system. The objective is to explore its performance for developing retweet prediction model.

- Compare the performance of proposed retweet prediction models with baseline models. The baseline models would be developed based on basic matrix factorization technique as well as machine learning technique using some commonly used explicit and implicit content based features. Furthermore, we want to compare the performance of retweet prediction models developed using machine learning technique and matrix factorization technique. The difference between machine learning and matrix factorization retweet prediction model is that the former predicts based on human extracted features whereas the latter predicts based on machine extracted latent features.

1.4 Proposed Approach

In this study, we propose to build a retweet prediction model which learns from the features related to users' behavior to predict their future retweet decision. Twitter users' data was downloaded using Twitter API, which includes their posts and followee information. We will then apply machine learning and matrix factorization techniques to build retweet prediction models and compare their performance in terms of predicting retweet decision. A user's behavioral patterns can be defined by their interests and attitudes. Implicit content-based features are extracted from users' posts to explain their interests and attitudes. Explicit features such as hashtags, URLs, and usernames are also included because of their successful use in past researches. Topic is included because it is a well-established important content-based implicit feature for developing retweet prediction model. Past researchers used general Latent Dirichlet Allocation (LDA) based topic extraction technique, which might not be suitable for short-length tweets. Therefore, we use twitter-LDA (Zhao et al., 2011) which was developed to extract single topic for short-length text. Instead of 2-dimensional sentiment as used in many past research works, we use 10-dimensional emotion-sentiment scores (Mohammad & Turney, 2013; Bravo-Marquez et al., 2016; Yarkoni, 2010; Plutchik, 2001), considering that the diversified and detailed representation of emotion has greater ability to define a user's behavior. The 10-dimensions include positive, negative sentiment along with 8-dimensional emotion such as anger, anticipation, joy, sadness, disgust, fear, trust, and surprise. 35-dimensional personality (Yarkoni, 2010) reflected by a user's tweets is also used to capture the effect of personality on a user's retweet decision. The proposed machine learning based prediction model calculates a user's past profile and finds the similarity between a target tweet and the user profile using different explicit and implicit content features. These similarities are then considered as features for retweet prediction model and fed to machine learning algorithm to predict retweet decision.

This research hypothesizes that a user's behavior is different as an author and retweeter and their past retweets provide more information than their past tweets in case of developing retweet prediction model. To validate this assumption, a user's profile is created in three different ways – using their past retweets only, tweets only, and tweets-plus-retweets. This work investigates

whether retweet-only profile provides better prediction accuracy than tweet-only or conventional tweet-plus-retweet-based profile.

For this research, matrix factorization technique is used to build retweet prediction model. Matrix factorization is a popular method to build recommender systems. This work examines the performance of matrix factorization model in case of predicting users' retweet decision. Matrix factorization technique explores the latent relations between users and messages for the purpose of making retweet decision. It has the ability to extract latent features of the items and the user's preference on these latent features, which may not be easily extractable by human experts. In case of matrix factorization models, user-message retweet matrix is factorized to learn user-message latent features for the purpose of retweet prediction. Newly introduced regularization term is used to include tweet features in the basic matrix factorization model, and is implemented using two different approaches. It calculates the similarity between a user's messages for the purpose of predicting user's retweet decision. These message-similarity-based new regularizers are developed assuming that, a user's retweets are similar to one another whereas the non-retweets are dissimilar to retweets and if messages are similar in the observed space, they will be similar in the latent space as well. Decreasing the difference between their similarity in the observed and latent space would help to discover more accurate user-message relation based on underlying latent features extracted by the factorization process. In case of message similarity calculation, their explicit and implicit content features are used. URL, hashtag, and user-mention are the explicit features. Topic, emotion, and personality are the implicit features.

Matrix factorization and machine learning are two common approaches used in the past for retweet prediction. This research would like to explore and compare their performance when extra information can be included (topic, emotion, and personality). For machine learning, the focus is more on feature set. This work is different from previous works on the feature set used. This work includes more features, explores different combinations of these features, and treats tweets and retweets separately. For matrix factorization, since plain matrix factorization models only consider retweet relation between user and tweet, extra information of tweets is not fully used. This research investigates the ways of including tweet features into matrix factorization model through regularization terms.

1.5 Assumption and Scope

Objective of this research is to build retweet prediction model which incorporates users' behavioral pattern for predicting their retweet decision. To calculate a user's emotion and personality as representation of their behavior, this work relied on previously published well accepted work (Mohammad & Turney, 2013; Bravo-Marquez et al., 2016; Yarkoni, 2010). Therefore, the accuracy of this part of the system is constrained by the accuracy of the word mapping correlation in the database they have developed. Word-emotion correlation database proposed by Mohammad and Turney (2013) and Bravo-Marquez et al. (2016) were used for emotion calculation. For personality calculation, this work used word-personality correlation database proposed by Yarkoni (2010). Mohammad and Turney (2013) and Yarkoni (2010) involved human subjects to develop their correlation database. Bravo-Marquez et al. (2016) extended the work done by Mohammad and Turney (2013). This research includes content-based features and does not include any social relationship features and structure-based features. Social relationship features and structure-based features will be explored in future research on user behavior analysis.

1.6 List of Publications to Date

- Firdaus SN, Ding C, Sadeghian A. Retweet prediction considering user's difference as an author and retweeter. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on 2016 Aug 18 (pp. 852-859). IEEE.
- Firdaus SN, Ding C, Sadeghian A. Topic specific emotion detection for retweet prediction. International Journal of Machine Learning and Cybernetics. 2018:1-3.
- Firdaus SN, Ding C, Sadeghian A. Retweet: A popular information diffusion mechanism–A survey paper. Online Social Networks and Media. 2018 Jun 30;6:26-40.

1.7 Organization of Chapters

- Chapter 2 presents some recent published research in the area of retweet prediction.
- Chapter 3 presents the methodology used to develop retweet prediction model. The architecture, feature extraction strategy, and prediction techniques are described in this chapter.
- Chapter 4 describes the experiment design and result analysis. It includes the description of dataset as well. The comparison of performance of different feature sets, different profile generation strategies, and different prediction techniques are also discussed in this chapter.
- Chapter 5 concludes the thesis work along with some proposals for future research.

Chapter 2

Related Work

In today's world, online social networks are considered as an accessible and fast medium to communicate and share information. The content shared on the social network can represent one's values and beliefs. On Twitter, through short messages known as tweets, users express their conscious as well as unconscious state of mind in a carefree manner. Information retrieved from social networks is excellent for predicting users' opinion, sentiment, taste, and interest in diversified areas such as politics, business, marketing etc. (Asur & Huberman, 2010; Boecking et al., 2015; Gupta et al., 2012; Liu et al., 2014; Mittal & Goel, 2012; Tumasjan et al., 201;, Zhao et al., 2014). Social network is a rich source of information to explore user's mental state regarding a topic, news, or product (Lim & Buntine, 2014; O'Connor et al., 2010; Ren & Wu, 2013; Roberts et al., 2012). Retweeting is a special activity of Twitter users, which allows them to repost the original tweets and thus act as a medium for information diffusion.

Retweet prediction is an imperative area of research due to its importance in understanding user's intention and approach in dispersing information (Jenders et al., 2013; Kim & Yoo, 2012; Kwak et al., 2010; Naveed et al., 2011; Pfitzner et al., 2012; Starbird & Palen, 2012; Starbird & Palen 2010). The broad spectrum of retweet prediction related research includes research on prediction of retweets (Huang et al., 2014; Macskassy & Michelson, 2011; Wang et al., 2015; Jiang et al., 2015; Vougioukas et al., 2017; Zhang et al., 2016), retweeters (Luo et al., 2013; Lee et al., 2015), retweet counts (Can et al., 2013) as well as tweet recommendation (Uysal & Croft 2011; Lu et al., 2012; Chen et al., 2012). Some research papers explore, analyze, and predict user's retweet activity, some papers are focused on finding out potential retweeters, whereas

other papers investigate the underlying reasons of why some tweets get more retweets or are spreading more virally. These research papers can be categorized based on their focus and research questions that they try to answer in their work.

We can define three main retweet related research questions as follows:

1. Which tweet will be retweeted by the user?
2. Who will retweet the target tweet?
3. Why do some tweets get more retweets?

In these papers, users' retweeting activity is mainly investigated from two perspectives: local and global. In case of local perspective, retweeting activity is explored from individual user's point of view. Every user's profile and interest are investigated to explore his retweet decision. The first research and second research question are focused on retweet activity from local perspective. In case of global perspective, tweets' general characteristics are investigated to find their retweetability. These types of research papers are focused on the third research question.

Retweeting activity is mainly dependent on three factors: user or reader of the tweet, author of the tweet, and content of the tweet. User represents the target user who gets the tweets in his timeline and decides the retweet action; author is the publisher of the target tweet; and content represents the target tweet itself including the words used in the tweet, their meanings as well as the overall information carried by the tweet. Every factor can be described by multiple features. The relation between user and author is one type of feature that is associated with both of them. We can consider it either as an author factor or a user factor. To make our discussion unambiguous, in the rest of this chapter, we treat it as a user factor. In this chapter, along with retweet prediction research, we have also discussed some research on tweet recommendation and retweeter prediction because these research works are quite related to retweet prediction. Tweet recommender system predicts retweets to build recommendation model considering retweet as an indicator of user's preference. Retweeter prediction explores user's interest on tweets to find potential retweeters.

2.1 Categorization of Research Papers

In Table 2.1, we have categorized the research papers based on the research questions they try to answer in their work. Though many research works have been done on retweet, their primary objectives can be different. In this section, we have described the categorization of retweet related research papers based on their primary objective to solve one of the three Twitter-specific research questions through their work.

The first research question is focused on investigating and predicting the tweets which will be retweeted by the user. These research papers can be further categorized based on their primary objectives. In the first sub-category, the primary objective is to analyze and investigate the factors that have influence on users' retweet activity. In these papers, researchers listed all features that might have impact on users' retweeting activity and then they analyzed the effects of these features on users' retweeting behavior to identify the most influencing features. Comarela et al. (2012) explored the effect from features such as user's prior interaction with author, author's tweeting rate, content of tweet on user's retweeting behavior. This research revealed some interesting behavioural details behind a user's retweet decision. Sun et al. (2013) studied the influence of serendipitous information on user's retweet behavior and showed that users like to propagate tweets containing serendipitous information. If a tweet is unexpected from a source as well as relevant to the user, then it is serendipitous.

In the second sub-category, the objective is to not only explore and analyze the features influencing user's retweeting behavior but also propose retweet prediction models based on their investigated features. Research papers in this spectrum, investigate and predict retweet behavior from the perspective of individual users. Peng et al. (2011) explored content influence, network influence, and temporal decay factor on users' retweeting decision and proposed Conditional Random Field (CRF) based retweet prediction model using features that define tweet's content influence, user's network influence, and temporal influence on user's retweet decision. Zhang J. et al. (2013) explored the influence of friends from a user's ego-network on their retweeting activity and then proposed retweet prediction model using only their explored features based on social influence locality. Zhang et al. (2015) explored influence of author, network structure,

content of tweet, and temporal information on users' retweeting probability and then proposed Hierarchical Dirichlet Process based retweet prediction model incorporating these features. Xu & Yang (2012) analyzed different features to develop retweet prediction model from the perspective of individual users. Their purpose was to investigate the importance of different author-based, social-relationship based, and content-based features on users' retweet decision. They explored the effectiveness of individual feature by developing and comparing the performance of retweet prediction models with different features. Yang et al. (2010) also analyzed different features related to user interest, content of tweet, and time on users' retweeting behavior and then proposed factor-graph-based retweet prediction model. Xu et al. (2012) analyzed the influence of social friends and breaking news on users' retweeting behavior and incorporated these influences in their proposed mixture latent topic retweet prediction model. Hoang and Lim (2013) analyzed three behavioural factors: topic virality, user virality and user susceptibility on users' retweet decision and proposed a tensor factorization retweet prediction model which represents retweets as three-dimensional tensors based on the mentioned factors. We can see that author influence, social influence or friends' influence, and content of the messages are some common factors which had been explored by many researchers. These research works made remarkable contribution to the field because they worked with different datasets and used different mechanisms to describe as well as analyze the effects of these factors to build efficient retweet prediction models. Zhao et al. (2018) proposed image retweet prediction model. To learn user preference for image tweets, they developed image retweet modeling (IRM) network based on attentional multi-faceted ranking method using textually guided neural network. Their proposed IRM network uses users' past image retweets along with their associated text as well as user's following relation to develop neural network based prediction model.

In the third sub-category, the primary objective is to develop retweet prediction models based on already known features. The focus for these research papers is on the design of effective prediction models. They use different machine learning methods to build novel and accurate retweet prediction models. Huang et al. (2014) proposed a novel methodology based on Bayes model to find users' interest in different categories and predict their retweet decision depending on the interest measurement. Macskassy and Michelson (2011) developed different

Table 2.1: Paper categorization based on the research questions

Research Question	Primary Objective	Title	Reference
Which tweets will be retweeted by user?	Analyze features influencing retweet activity	Understanding factors that affect response rates in Twitter	Comarela et al. (2012)
		Unexpected Relevance: An Empirical Study of Serendipity in Retweets	Sun et al. (2013)
	Analyze features influencing retweet activity and build retweet prediction model based on those features as well	Retweet modeling using conditional random fields	Peng et al. (2011)
		Social influence locality for modeling retweeting behaviors.	Zhang et al. (2013)
		Retweet Behavior Prediction Using Hierarchical Dirichlet Process	Zhang et al. (2015)
		Analyzing user retweet behavior on twitter	Xu and Yang (2012)
		Understanding retweeting behaviors in social networks	Yang et al. (2010)
		Modeling user posting behavior on social media	Xu et al. (2012)
		Retweeting: An act of viral users, susceptible users, or viral topics?	Hoang and Lim (2013)
		Textually Guided Ranking Network for Attentional Image Retweet Modeling	Zhao et al. (2018)
	Design effective retweet prediction model	Retweet behavior prediction in twitter	Huang et al. (2014)
		Why do people retweet? anti-homophily win the day!	Macskassy and Michelson (2011)
		A Multidimensional Nonnegative Matrix Factorization Model for Retweeting Behavior Prediction	Wang et al. (2015)
		Message clustering-based matrix factorization model for retweeting behavior prediction	Jiang et al. (2015)
		Identifying retweetable tweets with a personalized global classifier	Vougioukas et al. (2017)
		Retweet prediction with attention-based deep neural network	Zhang et al. (2016)
	Design effective tweet recommendation model	User oriented tweet ranking: a filtering approach to microblogs	Uysal and Croft (2011)
		Twitter User Modeling and Tweets Recommendation Based on Wikipedia Concept Graph	Lu et al. (2012)
		Collaborative personalized tweet recommendation	Chen et al. (2012)
		Learning to Rank Tweets with Author-based Long Short-Term Memory Networks	Piao and Breslin (2018)
Who will retweet the target tweet?	Finding out retweeters	Who will retweet me? Finding retweeters in Twitter	Luo et al. (2013)
		Who will retweet this? Detecting strangers from Twitter to retweet information	Lee et al. (2015)
Why do some tweets get more retweets?	Finding out the reasons behind spreading of information by retweet activity	RT to Win! Predicting Message Propagation in Twitter	Petrovic et al. (2011)
		Want to be retweeted? large scale analytics on factors impacting retweet in twitter network	Suh et al. (2010)
		Predicting retweet count using visual cues	Can et al. (2013)
		Modeling and predicting retweeting dynamics on microblogging platforms	Gao et al. (2015)
		Bad news travel fast: A content-based analysis of interestingness on twitter	Naveed et al. (2011)
		Analyzing and predicting viral tweets	Jenders et al. (2013)
		Emotional divergence influences information spreading in Twitter	Pfitzner et al. (2012)
		Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior	Stieglitz and Dang-Xuan (2012)
		Role of sentiment in message propagation: Reply vs. retweet behavior in political communication	Kim and Yoo (2012)
		Assessing the reTweet proneness of tweets: predictive models for retweeting	Nesi et al.(2018)

retweet prediction models: general/random decision based, recent communication based, on-topic based, and homophily based, to have detailed understanding on users' retweet decision. Wang et al. (2015) and Jiang et al. (2015) proposed matrix factorization retweet prediction models. Wang et al. (2015) used user-based and content-based features to incorporate user similarity, activity, interest, and content's influence on their retweeting activity and developed nonnegative matrix factorization retweet prediction model. Jiang et al. (2015) tried to avoid the complexity of finding user similarity in a large network. So, they only utilized the impact of message similarity on users' retweeting behavior and proposed message-clustering-based retweet prediction model using matrix factorization technique. Zhang et al. (2016) designed retweet prediction model using attention based deep neural network incorporating users' interests and user/author information. The capability of deep neural network to learn optimal features automatically helped them build a state-of-the-art prediction model without the complex task of feature engineering. Vougiouk et al. (2017) investigated the effectiveness of different feature sets based on user, author, and content of the tweet, built logistic-regression-based personalized retweet prediction model and proposed a state-of-the-art model with only 10 features. These research works are mainly focused on the design of prediction models and they did not intend to analyze the influencing features on users' retweet behavior, rather they explored different machine learning techniques to design one competent model.

In the fourth sub-category, researchers worked on the detection of retweetable tweets in order to design tweet recommender system considering retweetable tweets as users' preferred item. Uysal and Croft (2011) explored different user, author, and content-based features to define a tweet's interestingness and used learning to rank strategy to define retweet-likelihood-based tweet ranking. Lu et al. (2012) considered retweets as tweets relevant to users' interest and ranked tweets based on their similarity with user profile developed using Wikipedia concept graph. Assuming users' retweeting action as their personal preferences based on usefulness and informativeness of the tweets, Chen et al. (2012) developed a personalized tweet recommender system using collaborative ranking method. Though primary objective of these research papers is to design personalized tweet recommender system, they did investigate on the behavior of the retweetable tweets because retweetability is a good indication of being a good recommendation candidate. Piao and Breslin (2018) proposed deep neural network based tweet recommender

system. They used author-based long short-term memory to learn the latent representations of tweets. Prediction of prefereed tweets was done based on the similarity between the author and the tweet along with author's similarity with the user.

The second research question is related to finding potential retweeters, or identifying who is more likely to retweet a tweet among all the followers of the author of the tweet. Since retweet is a significant mechanism for information diffusion, finding out proper target users is an important task in order to spread the information efficiently. Luo et al. (2013) and Lee et al. (2015) both were focused on prediction of potential retweeters for target tweet though their approaches towards the problem are different. Lee et al. (2015) aimed to find out retweeters among users who are requested to retweet the tweets on a specific topic. Their purpose was to find out potential retweeters to spread information during an emergency case. Luo et al. (2013) used learning to rank model to rank followers based on their retweet probability for a target tweet. In case of finding out potential retweeters, researchers mainly put emphasis on feature sets that define followers' intentions and activities for the task of retweeting.

It was observed that some tweets have the potential to be retweeted more by the users. Researchers focusing on the third research question explored the underlying reasons that caused the virality of tweets. These research papers did not predict or study retweet behavior from the perspective of individual users, rather they explored the retweetability of a tweet from global perspective. Suh et al. (2010) explored a large number of content-based and contextual features to find their underlying association with the tweet's retweetability. The objective of Petrovic et al. (2011) was similar to the work of Suh et al. (2010), but they explored relatively small number of tweets' content-based features and social features related to authors to predict retweetability of streaming tweets. Finding out effective features from tweets is a challenging task due to their length restriction. To overcome this limitation, Can et al. (2013) used visual cues from the image linked to the tweet to find its retweetability and showed that visual cues served as a competent added factor to find a tweet's retweet count. Gao et al. (2015) included the impact of tweets' age and users' time-dependent activity to find the popularity of tweets. They showed that not only the interestingness of tweets but their posting times also have effect on popularity of tweets. Researchers also investigated the impact of sentiments on tweets' retweet probability (Naveed et

al., 2011; Pfitzner et al., 2012; Stieglitz & Dang-Xuan, 2012; Jenders et al., 2013; Kim & Yoo, 2012). With different objectives, different settings, and different datasets, all these research papers found that tweets reflecting negative sentiments have higher probability to be retweeted by users. Nesi et al. (2018) used different features to predict retweet count of tweets. They included features related to tweet, author, and author's follow network. They identified publication time and listed count as relevant features along with some previously used features. They also showed that CART decision tree model with recursive partitioning procedure gave superior performance in predicting retweet count when compared with other machine learning models such as Random Forest and Stochastic Gradient Boosting.

2.2 Analysis of Retweeting Behavior

Users are the main actors in online social networks. They create, initiate and propagate information. So, users' retweeting behavior has been investigated and analyzed broadly in this area. Though a user's retweeting decision is subjective, results from these analyses can help us gain a better understanding on why they make these decisions. Comarela et al. (2012) showed that newer tweets, tweets from previously retweeted authors, and authors with lower posting rate have higher probability to be retweeted by the users. The study also showed that users like to retweet shorter tweets. The reason can be that, in case of shorter tweets, users might get room to add their own text. Zhang J. et al. (2013) investigated neighbours' influence on users' retweet activity. The experiment showed that a user's retweeting probability was positively correlated with number of their active friends whereas the probability was negatively correlated with the number of connected circles formed by those active friends. The reason can be that a user might not be interested to retweet a message which is already known by many of their neighbours. Zhang et al. (2015) and Petrovic et al. (2011) showed that, author of tweet has good influence on a user's retweet activity. When same microblog was posted by two different authors at different time slot, many users repost the microblog posted at earlier time even though another same post appeared first in their main page, which clearly indicates the influence of author on users' retweeting behavior (Zhang et al., 2015). It was found that an author's authority such as number

of followees, number of times the author was listed, and inclusion of teen-related topics increases the retweetability of a tweet (Petrovic et al., 2011).

A large-scaled analysis has been done to find features that have good impact on tweet's retweetability (Suh et al., 2010). Suh et al. (2010) used Principal Component Analysis (PCA) to explore influencing features and built Generalized Linear Model to explain the influence of these features on finding the retweet probability. According to their result, the number of followers and followees and age of the account of the author have positive influence on the retweetability of a tweet. On the other hand, there is no strong correlation between an author's total number of past tweets and their retweet rate. As per their analysis, hashtags and URLs have strong correlation with retweetability of a tweet and in case of URLs, the retweet rate varies in different domains. Sun et al. (2013) made an interesting finding that users like to diffuse serendipitous information. They defined serendipity as unexpected tweet from source (author) which is useful or relevant for receiver (user). They developed method using Likelihood Ratio Test to check unexpectedness and relevance of tweets. The unexpectedness test was eventually a test to find out whether the tweet can be explained by the perceived model (based on the received information from the source) of the source (developed by the receiver) or can be explained by the mixture model of multiple contexts. They also developed a preference model of the receiver based on his posts. Then they checked if the tweet is relevant to the user's posting. From this work, the researchers found that 27% of retweets in Twitter and 30% of retweets in Weibo contain serendipitous information.

Lee et al. (2015) built models based on users' personality traits, social behavior, social relations, and content of the tweets to see the willingness of the user to propagate information when they were asked to do so during the emergency case. In this research, a good number of features have been used to define a user as a potential retweeter or non-retweeter. Users' activity, personality, readiness (to retweet), and past retweeting behavior related features showed strong impact on classifying the user as retweeter. Researchers also confirmed that aging of a message has impact on its popularity to be retweeted. As per Gao et al. (2015), popularity of a message to be retweeted follows power law distribution with its aging process.

Many researchers investigated the impact of sentiments of tweets on users' retweeting behavior (Naveed et al., 2011; Pfitzner et al., 2012; Stieglitz & Dang-Xuan, 2012; Jenders et al., 2013; Kim & Yoo, 2012). Based on their findings, in general, tweets with negative sentiment were retweeted more by the users. But Jenders et al. (2013) showed that tweets with excessive negative sentiments do not have the potential to be viral. Stieglitz and Dang-Xuan (2012) investigated that in case of political information diffusion, messages containing positive or negative sentiment had higher probability to be retweeted by others. In this case as well, massages with negative sentiment were retweeted more than messages with positive sentiment. As a measure of sentiment, researchers mainly considered positive, negative, and neutral sentimental score of the tweet. Some researchers (Pfitzner et al., 2012, Jenders et al., 2013) also calculated emotional divergence of a tweet which is basically the normalized absolute difference between the positive and negative sentiment score of the tweet. Pfitzner et al. (2012) showed that highly emotionally diverse tweets had five times higher chance to be retweeted by the users. Naveed et al. (2011) used dictionary-based approach (Kim et al., 2009) to find sentiments of the tweets. As a measure of sentiment, they used valence, dominance and arousal score of a tweet. Researchers also used LIWC (Linguistic Inquiry and Word Count) program (Pennebaker et al., 2001) to find the sentiments of tweets based on the number of words in the tweet which belong to the following two LIWC categories: "Positive emotion" and "Negative emotion" (Stieglitz & Dang-Xuan, 2012). According to an experiment by Jonah Berger (Berger, 2011), an expert in viral marketing and social influence, people in high arousal state (after running or jogging) tend to spread information more than people in low arousal state (sitting still). Berger also showed that arousal always increases social transmission no matter it is positive (amusement) or negative (anxiety). Results of Burger's experiment somehow correlate with user behavior analysis research for retweeting, as it is found that users usually like to spread information containing non-neutral sentiment, especially negative sentiment.

2.3 Retweet Prediction

The research in retweet prediction is mainly conducted in four steps. In the first step, researchers collect Twitter dataset and then in the second step, various features belonging to

three factors (author, user, content) are extracted from the dataset. The third step includes design of retweet prediction model using the extracted features. The final step is to evaluate the proposed model. In this section, we discuss each of these steps and how they are implemented in different research works, as well as how the three tweet-related factors and their corresponding features are utilized for retweet prediction. Here we also discuss a few research papers on tweet recommendation because these works are focused on the same research question as retweet prediction (see Table 2.1) and follow the similar steps as the work on retweet prediction. We have also included a few research papers on retweeter prediction as they explore some important user-author relations and tweet content features to find users who might have interest to retweet the tweet.

2.3.1 Data Collection

For retweet prediction, datasets are not usually publicly available for research. Though a few research works used previously used dataset (Naveed et al., 2011; Jiang et al., 2015; Wang et al., 2015), most of the time researchers had to collect their own data. Researchers used Twitter Application Programming Interfaces (API) to collect data directly from Twitter (Uysal & Croft 2011; Peng et al., 2011; Lu et al., 2012; Xu et al., 2012; Zhang et al., 2016; Suh et al., 2010; Petrovic et al., 2011; Sun et al., 2013; Luo et al., 2013; Lee et al., 2015; Can et al., 2013). Twitter API allows users to download data based on the type and requirement of their research. Researchers used REST API when they needed to collect historical data based on some parameters and used streaming API when they needed real time data. These APIs also have some limitations, such as REST API allows to get at most 3,200 latest tweets from a user[1]. In case of searching tweets based on a query, standard version of Twitter REST API returns a sampling of recent tweets published in the last 7 days[2]. For getting real-time tweets using streaming API, standard version of API allows to track at most 400 keywords, 5000 users, and 25 locations[3]. Twitter provides enterprise versions of these APIs which allow the users to get

[1]https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

[2]https://developer.twitter.com/en/docs/tweets/search/overview

[3]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

elevated access to their data. But enterprise versions are not free while the standard ones are. Twitter also puts rate limit per request on getting their data. All necessary information of using Twitter APIs is available in their developer platform website[4].

Third party libraries such as Twitter4J[5], tweepy[6], twitter-python[7]can also be used to collect and process data from Twitter. Snowball sampling method was used to get information of large connected network (Macskassy, 2012; Macskassy & Michelson, 2011). In this method, researchers select some seed users and then collect data from users who are connected to the seed users (through retweet/mention) and this process continues until an adequate amount of data is obtained.

2.3.2 Feature Extraction

The accuracy of retweet prediction greatly depends on which features are used and whether they are effective in terms of predicting retweet. The past research has shown that author, user, and content of the tweet have great impact on users' retweeting decision. These factors could capture or reflect the impact of authors' influence, author and user social relations, users' interest, and content of the tweet on retweeting activity. Different features based on these three factors and their objectives are given in Table 2.2. Below is the description of features based on these factors.

2.3.2.1 Author of the Tweet

Intuitively it can be said that author of a tweet has good impact on its retweetability. Findings from the past research also support this intuitive observation. According to the study conducted by Cha et al. (2010), if a tweet is from content aggregation service or news media, or

[4]https://developer.twitter.com/en.html

[5]http://twitter4j.org/en/

[6]http://www.tweepy.org/

[7]https://github.com/bear/python-twitter/

from a popular and most mentioned user such as celebrities, it will get more retweets. Number of followers and followees of the author, age of the account, number of tweets from the author, tweet frequency (per day) of the author, number of tweets favoured by others, language of the author, ratios of retweeted tweets, ratios of tweets receiving replies, and whether the author is a verified user or local elite, are good features that can be used to measure an author's influence on the retweet decision (Suh et al., 2010; Petrovic et al., 2011; Uysal & Croft 2011; Xu & Yang 2012; Jenders et al., 2013; Chen et al., 2012).

2.3.2.2 User of the Tweet

One of the basic questions in retweet related research is "Which tweets will be retweeted by user?". From this research question, it is evident that user is the primary actor in retweeting activity. Since retweet is a personal decision, it is hard to find any definite answer to this question as the reasons for retweeting could be purely subjective and thus varied from user to user. The most common reasons could be listed as follows: the user wants to spread the information; the user finds it interesting enough to share with others; the user finds the tweet helpful for others; the user's relation with the author of the tweet influences him; the user is influenced by his neighbours in the social network.

Many features related to users have been explored for the purpose of retweet prediction, retweeter prediction, and tweet recommendation. Features that are used to measure the user-author relation include a user's recent communication with the author, interest similarity with the author, social relation with the author, and whether user is mentioned in the tweet (Macskassy & Michelson, 2011; Uysal & Croft, 2011; Chen et al., 2012). A user's interest profile can be derived from his past posts. Commonly used profiles include bag-of-word profile using the Term Frequency-Inverse Document Frequency (TF-IDF) weights of the words (Luo et al., 2013; Xu & Yang, 2012; Xu et al., 2012), hashtag-based profile (Xu & Yang, 2012), and entity-based profile (Xu & Yang, 2012). TF-IDF technique finds scores/weights for terms in user's tweets based on their importance in distinguishing the user from others. Thus, TF-IDF based profile has the capability to represent users uniquely. In case of hashtag and entity-based profiles, only distribution of hashtags and entities might not give much information because many users might use the same hashtag and entity. So, the preferred method is to check their weights (frequency)

while creating user profile because frequency of using hashtags and entities might give more information about a user's preference/interest. Another constraint of using entity-based profile is to select efficient method to extract entities from the tweets. Performance of entity-based profile is quite dependent on the efficiency and accuracy of entity extraction methods. Researchers mainly used AlchemyAPI[8] to extract entities from user's tweets. Some research works use third-party knowledge base to create the user interest profile. Macskassy and Michelson (2011) used Wikipedia's knowledge base to create user's topic of interest profile. They identified the entities from user's tweets and categorized them based on their category in Wikipedia page. They also matched the content of entities in this process to solve the problem of ambiguity. The categories of mentioned entities were used to define user interest profile.

According to the research (Yang et al., 2010; Wang et al., 2015), oftentimes, users' retweet decisions are consistent with the similarity degree between the user profile and the content of the tweets. Different similarity measures such as cosine similarity and Jaccard similarity have been used to calculate the similarity between user profile and content of the tweet. The calculated similarity scores are then used as potential features for retweet prediction (Xu et al., 2012; Xu & Yang, 2012; Yang et al., 2010; Chen et al., 2012; Vougioukas et al., 2017). In case of cosine similarity measure, user profile and target tweet profile are defined by the vectors represented by the distribution of topics/terms/hashtags in user profile and in the target tweet respectively. According to this similarity measure, user profile and target tweet profile are similar if these vectors share similar orientation and the angle between them is small. Jaccard similarity measure was used when the researchers define user interest as a set of terms derived from his past posts and target tweet was defined by the set of terms used in the target tweet.

Friends' influence on user can also be used as predictive features. Zhang J. et al. (2013) used data from Weibo micro-blogging service[9]. They defined social influence locality as a function to measure how a user's retweet decision is influenced by their active neighbours (users who have already retweeted the target tweet). The designed social influence locality function was

[8]http://www.alchemyapi.com/

[9]Chinese micro-blogging service, allows its users to repost the tweets similar to Twitter's retweeting service

based on pair-wise influence and structure influence. In case of pair-wise influence, they used Random Walk with Restart (RWS) method to calculate random walk probability for each active neighbour of the given user to reach the given user following the network connection. In case of structural influence, they used a linear combination of the number of connected circles formed by the active neighbours.

In case of finding retweeters, influencing features are mainly related to a user's activity, intention, and interest. Lee et al. 2015 explored a large number of features in this regard. They defined a user's activeness by features such as average retweets per day, tweeting likelihood of the day (hour), tweeting steadiness, number of status messages (Lee et al., 2015). They also included personality scores derived from user's posts to describe the impact of user's personality on his retweeting activity.

2.3.2.3 Content of the Tweet

Both explicit and implicit features related to the content of the tweet are used in retweet prediction models. Some of the explicit or directly measurable features are presence/absence of hashtag, URLs, emoticons, positive-negative words, punctuation marks, username, first person pronoun, second person pronoun, third person pronoun (Naveed et al., 2011; Uysal & Croft 2011; Kim & Yoo 2012). Researchers also used language and length of the tweet as features for prediction model (Petrovic et al., 2011; Chen et al., 2012, Uysal & Croft 2011). Popular implicit content-based features related to tweet include topics, terms with their TF-IDF scores, topic novelty, topic virality, sentiment, emotional divergence, etc. (Naveed et al., 2011, Uysal & Croft 2011; Petrovic et al., 2011; Xu & Yang, 2012; Yang et al., 2010; Pfitzner et al., 2012; Jenders et al., 2013; Hoang & Lim, 2013; Chen et al., 2012).

Some of the algorithms and tools used to extract implicit features include Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Term Frequency-Inverse Document Frequency (TF-IDF) (Leskovec et al., 2014), Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001),

SentiStrength (Thelwall et al., 2010), and AlchemyAPI[10]. LDA is used to determine a user's topics of interest or topics of tweets. It is a generative statistical method that considers each document as a collection of topics and finds the latent topics in the document (Blei et al., 2003; Blei, 2012). TF-IDF is a statistical measure used to find out the importance of a word for a document in a collection of documents. In this method, the importance (or weight) of a word increases proportionally by its number of occurrences in the document but is counterbalanced by its frequency in the whole corpus. This measure has been used by the researchers to generate user's bag-of-word profile which consists of his preferred words based on their TF-IDF weights. This profile can represent user's content-based interest. Researchers use LIWC[11] technique to find different text-based features. It finds the percentage of words in a document/text which belong to more than 70 different categories. These categories include simple linguistic factors such as Word Count, first person pronoun, as well as factors which indicate affect and emotion such as positive or negative emotion. AlchemyAPI uses machine learning technique to perform text analysis tasks. It is used to find entities in a user's tweets (Xu & Yang, 2012; Xu et al., 2012). Sentiments of tweets can be determined by SentiStrength method. SentiStrength is a lexicon-based approach which uses linguistic rules to find the positive and negative sentiment score of a tweet. Researchers also use LIWC (Stieglitz & Dang-Xuan, 2012) and dictionary-based approach (Naveed et al., 2011) to find sentiments of tweets. Affective Norms of English Words (ANEW) is a dictionary (Bradley & Lang, 1999) that gives numerical values of 1,030 words for three attributes indicating emotions: valence, dominance, and arousal. Valence refers to the degree of goodness/pleasantness (from displeasure to pleasure) invoked by the word, dominance refers to the extent of dominance (from weakness to strength) denoted by the word, and arousal refers to the degree of arousal (from calmness to excitement) evoked by the word. The total values of these three attributes for a tweet were the summation of these values for each word in the tweet.

[10]https://www.ibm.com/watson/alchemy-api.html

[11]http://liwc.wpengine.com/

Though different approaches have been used to define sentiment of tweets, accurate detection of sentiments in tweets can be a little tricky because of the presence of informal words in tweets. Dictionary that is developed particularly for Twitter can potentially solve this problem. For example, SentiStrength is considered as a better detector than other general dictionary-based methods because it was developed to find sentiments from short informal text. Bravo-Marquez et al. (2016) also extended general word-emotion lexicon developed by Mohammad and Turney (2013) to include informal words used in Twitter and we used (Firdaus et al., 2017) this lexicon to extract emotion-sentiment from tweets and achieved good results.

2.3.3 Prediction Model

Retweet prediction models are normally developed using different machine learning techniques. The fundamental task is to utilize the extracted features to develop effective retweet prediction model. It could be considered as a typical classification task. So, the basic retweet prediction model consists of feature extraction step followed by machine learning step to classify a tweet as "to be retweeted" or "not to be retweeted" based on the extracted features. It is important to find right features that are useful for the prediction task and to find right learning algorithms to make accurate predictions. In this section, we discuss some strategies to build retweet prediction models.

Many retweet prediction models have been developed based on the aforementioned two-step process and usually the original machine learning algorithms are used as they are. Zhang J.et al. (2013) defined functions to model social influence locality features. Social influence locality implies that user's retweeting behavior is influenced by close friends in the ego-network. They developed logistic regression classifier to build their prediction model using social influence locality features. Xu and Yang (2012) developed a retweet prediction model where they created TF-IDF, LDA, hashtags, and entity-based user profile. Cosine similarities between user profile and the target tweets were used as content-based features of their prediction model. Using some author-user relation features, content-based features, and author-based features, they developed three different prediction models using three machine learning techniques: decision tree, support

Table 2.2: Features and their objectives based on different factors

Factors	Objective	Related features
Author of the tweet	To define author's (global) influence on his tweet's retweetability	Author's number of followers and followees, age of author's account, number of tweets from the author, tweet frequency (per day) of the author, author's number of tweets favoured by others, language of the author, ratios of author's tweets retweeted by others, ratios of author's tweets received replies, and whether the author is a verified user or local elite
User of the tweet	To include user-author relation on a tweet's retweetability	User's recent communication with the author, user's social relation with the author, user mentioned in the tweet, author's influence on user (friend's influence)
	To define user-author interest similarity on a tweet's retweetability	User's interest similarity with author's interest.
	To include impact of user's activity on his retweeting probability	User's average retweet per day, tweeting likelihood of the day (hour), tweeting steadiness, and number of status messages, tweeting likelihood of the hour (to find retweeters), URLs/hashtags/mentions per day in posts
	To include the impact of user's personality on his retweeting activity	User's Big 5 and their 30-sub dimensional personality scores based on his posts
Content of the tweet	To include the impact of term distribution in tweet on its retweetability.	Presence/absence of hashtag, URLs, emoticons, positive/negative words, punctuation marks, username, first person pronoun, second person pronoun, third person pronoun, language, length of tweet.
	To define the impact of tweet's topic on its retweetability	Topic of the tweet, novelty and virality of tweet's topic
	To define the impact of tweet's terms on its retweetability	Importance of terms in the tweet (using TF-IDF scores), hashtags and URLs in the tweets
	To include tweet's emotion/sentiment on its retweetability	Emotion/sentiment reflected by the tweet, emotional divergence indicated by the tweet

vector machine, and logistic regression. Vougioukas et al. (2017) explored a wide range of author-based, user-based, and content-based features and used logistic regression method to build retweet prediction model. Through experiment, they identified 10 most effective features to create the prediction model. Can et al. (2013) proposed retweet count model based on visual cues of an image that is linked to the tweet. Along with two low-level features such as color histograms (distribution of color intensities in the image) and GIST (set of perceptual dimensions), the researchers used object-based feature (Li et al., 2010). Object-based feature was a set of object detectors to detect 177 objects in the image. They used 3 different regression methods: linear, SVM, and random forest to built their retweet count model using different features. Some research works design retweet prediction models based on their own prediction strategies and novel models are proposed. Macskassy and Michelson (2011) built four models to find the probability of a tweet to be retweeted by a specific user. The first model is called general model in which a user will randomly retweet a tweet with higher probability on more recently seen tweet. The second model is recent communication model in which a user will retweet tweets from authors with whom he has recent communications. The third model is called on-topic model in which the probability of a tweet to be retweeted is high if its profile is similar to user's interest. And the last model is homophily model in which a user will retweet tweets from authors with similar taste. The objective of this research was to find out the most effective model which can predict a user's probability to retweet a tweet. On a dataset with 79k tweets, the proposed homophily model showed best performance followed by recency model, on-topic model, and finally general model. They also found that a user's retweet behavior is better predicted by multiple models instead of one. The retweet prediction model proposed by Huang et al. (2014) measures a user's interest in following categories: technology, politics, life, sports, entertainment, health, travel, and finance. Then for retweet prediction, they computed the probability of the target tweet belonging to a final category; if this probability is greater than user's interest in that category, they predicted that the user is going to retweet the target tweet.

Researchers have also adapted and modified existing machine learning methods to make them more fit as retweet prediction models. Petrovic et al. (2011) used different author-based and content-based features to design their model. They used Passive-Aggressive algorithm (PA) (Crammer et al., 2006) based machine learning approach to design a model to predict streaming

retweets. They customized the original prediction rule of PA algorithm to adapt the time-sensitive rules for retweeting (e.g., tweets containing a specific word might have higher probability to be retweeted in the morning than in the evening). Zhang et al. (2015) adapted Hierarchical Dirichlet Process (HDP) (The et al., 2012) to design a nonparametric statistical method for retweet prediction. They incorporated structural, textual, and temporal information in their proposed HDP model. First, they extended HDP to model author, structure and content information. In the model, for each followee, the probability of retweeting his posts was subjected to binomial distribution with beta error. Structural influence (influence from neighbours) for users was also modeled by Beta distribution. Content influence was modeled by hidden topics and HDP-based generative process finds the topic assignment of microblogs. In the retweet prediction phase, the weights of recent topics were increased to incorporate temporal information and the retweeting probability of microblogs was then calculated. Peng et al., (2011) proposed Condition Random Field (CRF) based retweet prediction model. Assuming user's retweet decision is influenced by local and network factors, they chose conditional random field to find the retweet probability conditioned on features related to the target tweet and target user. The researchers were concerned about the conditional distribution of user decision given the new tweet and the user. In their proposed method, they modeled tweet's content influence as well as network influence on user's retweet decision. For content influence, they included similarity between tweet's content and user's interest, similarity between tweet's content and user's friends' or followees' interest, and similarity between global interest (determined based on all tweets and retweets in the dataset) and tweet. They also included URLs, hashtags, and mention-based features to model tweet's content influence on user. To define network influence, they used author-based features such as author's number of followers/followees, author's number of tweets/retweets; and author-user relationship based on common followers, followees, mentions, and retweets. They utilized retweet network's "small world" (Watts & Strogatz, 1998; Adamic, 1999) nature to design an efficient graph partitioning algorithm to make their method suitable for large, complicated network. In case of small world network, the network graph is highly clustered, average path length (APL) between all pairs of nodes is small, and an individual is mainly influenced by a small number of his connections. Retweet network can be considered as small world network because retweets are spreading through the connections of users and these

connections normally show clustering property. In Twitter, the APL between pairs of nodes is small, and retweet network can be defined by fraction of edges (portion of connections) which make the clustering structure of the network. Xu et al. (2012) proposed a mixture latent topic model to explore user's retweet behavior. Assuming user's posting behavior is influenced by breaking news, posts from friends, and his intrinsic interest, the researchers extended the widely used author-topic model (Rosen-Zvi et al., 2004) to include the mentioned factors to build their proposed mixture topic model. Yang et al. (2010) proposed a semi-supervised factor graph model to predict users' retweeting behavior based on factors such as user, message, and time.

Matrix factorization is an effective technique used by the researchers to design retweet prediction model. The fundamental task of matrix factorization technique is to factorize the observed user-message retweeting matrix $R \in \mathrm{R}^{MxN}$ for M users and N messages into two low dimensional matrices $P \in Mxk$ and $Q \in kxN$ such that product of P and Q approximates R. The main objective is to find the latent features k which defines the latent relationship between user and message. Jiang et al. (2015) proposed message-clustering-based matrix-factorization models assuming that if messages are similar in observed space then they are similar in latent space as well. So, they extended the basic matrix factorization model by using clustering-based regularization term. Different content-based features were used to find the similarity between messages which was then used to define cluster of messages. Wang et al. (2015) proposed two matrix-factorization-based retweet prediction models. They used strength of social relationship between users to generate objective function for user-based prediction model. Another prediction model was developed using content-based features. Finally, they fused both models based on their error rates. Hoang and Lim (2013) represented retweets as three-dimensional tensors of authors, their followers, and tweets themselves. Then they proposed a tensor factorization model to derive three behavioural factors - topic-specific user virality, topic-specific user susceptibility, and topic virality. These factors were then used as features to predict user retweet actions.

Nowadays deep learning methods become popular for their efficiency and ability to learn optimal features automatically. Zhang et al. (2016) proposed a retweet prediction model using attention-based deep neural network. In this model, they used convolutional neural network to encode content of the tweet and attention-based neural network to encode the attention interest of

the user. Similarity between a user's attention interest and a tweet was also computed. They encoded each user and author with continuous vector. Finally, a concatenation layer was used to produce a hidden state using these vectors and a fully connected Softmax function was used for retweet prediction.

Researchers working with tweet recommender systems consider retweet as a mechanism to identify a user's preference. These research works predict retweets to check which tweets are retweeted or preferred by the user. Uysal and Croft (2011) explored users' retweeting behavior to filter tweets for individual users. They used author-based, user-based, and content-based features to develop a decision-tree-based classifier to classify tweets as retweetable or not for a specific user. They used learning to rank method to rank incoming tweets to develop tweet recommendation list for a user. Lu et al. (2012) built a tweet recommender system by ranking incoming tweets based on their similarity with user profile. In this research, a user's retweets are considered as relevant tweets for recommendation. Their novel approach to create user profile using Wikipedia concept graph showed better performance for tweet recommendation compared to models with TF-IDF based profile. Chen et al. (2012) developed a personalized tweet recommendation method assuming retweets as a measure of a user's interest and authors of retweets as a measure of social relationship. They included topic level user interest and user-author relation features to build collaborative-ranking based tweet recommender system.

To design model to predict retweeters, Luo et al. (2013) used SVM^{Rank} method to rank potential retweeters based on their probability to retweet. Lee et al. (2015) explored a wide range of features and built different prediction models to compare their results using the following machine learning techniques: Random Forest, Naïve Bayes, Logistic, SMO, and AdaBoostM1. They found that random forest model performed best in predicting potential retweeters.

Retweet prediction would become more accurate if data from active users are used to learn and train the model. Social networks have many active as well as inactive users. Inclusion of inactive users' data might not have accurate contribution in prediction model. So, finding out active users is an important step when collecting the data. On the other hand, most of the time, retweet prediction model needs both positive and negative examples. The positive examples are a person's retweets. Negative examples are the tweets which are posted by the user's followees and

appear in the user's timeline, but not retweeted by the user. The reasons for which a user is retweeting a target tweet is somehow understandable and derivable. However, the reasons for which a target tweet is not retweeted by the user is tricky to find out and they often are decided by many unseen features. We cannot just say that the user did not like the tweet or the user is not interested in the topic of the tweet. A user might not retweet a target tweet for some concealed reasons such as he might not see the tweet, he might not be active during the posting time of the tweet, he might not be in the frame of mind to spread any information. Researchers handled these issues in different manners. Zhang J. et al. (2013) predefined 6 timestamps to define negative instances. If a tweet is not retweeted by the user within any of the mentioned timestamp (selected randomly), then they considered it as negative instance. Zaman et al. (2010) used one-hour time window to see if a tweet is retweeted by the user within one hour of its posting time. If the tweet is not retweeted by the target user within one hour, then that tweet was considered as negative instance. Uysal and Croft (2011) selected active users based on the following three criteria: he has 10-1000 friends/followers, tweets 1~200 times a week, and tweets more than 10 times. By considering only the active users they eliminated the uncertainty to some extent about the user not seeing the non-retweeted tweet. Xu and Yang (2012) first selected seed/active users who have 100-3000 followers and followees, are listed 1~50 times and have 10~200 tweets per week.

2.4 Evaluation

The last important step of retweet prediction is to evaluate the performance of the model. In case of prediction, dataset is divided into training and testing set; the model is trained using training dataset and tested using testing dataset. Machine learning techniques analyze the training data (instances with observed outcomes) and learn reasoning to find the outcome for the instance. Testing dataset (instances with unknown outcome) is used to find performance of model on unseen data. Standard approach is to use 70%-90% data as training samples and the rest for testing. Another popular approach to evaluate the learning model is *k*-fold cross validation. In this technique, the dataset is divided into *k* equal subsets then *k-1* subsets are used as training data and the remaining subset is used as testing data. This process is repeated *k* times

Table 2.3: Evaluation metrics used in different research

Research	Precision	Recall	F1-score	Accuracy	AUC	MAP	AUPRC	RMSE	ROC
Peng et al. (2011)	x	x	x						
Zhang J et al. (2013)	x	x	x	x					
Zhang et al. (2015)	x	x	x						
Xu and Yang (2012)	x	x	x						
Yang et al. (2010)	x	x	x						
Xu et al. (2012)	x	x							
Hoang and Lim. (2013)							x		
Huang et al. (2014)	x								
Wang et al. (2015)	x	x	x						
Jiang et al. (2015)	x	x	x	x					
Vougioukaset al. (2017)	x	x	x						
Zhang et al. (2016)	x	X	x						
Uysal and Croft (2011)	x	X	x						
Chen et al. (2012)						x			
Luo et al. (2013)						x			
Lee et al. (2015)			x		x				
Petrovic et al. (2011)			x						
Can et al. (2013)								x	
Naveed et al. (2011)									x
Nesi et al. (2018)			x						
Piao and Breslin (2018)						x			
Zhao et al. (2018)					x				

such that every subset is used exactly once as test data. Finally, the results from all iterations are averaged to get the final result. Performance of the learning model was evaluated using several metrics. Many evaluation metrics are available; researchers picked the one suitable for their work. In case of predictive model, researchers usually picked the following metrics: accuracy, precision, recall, and F1-score. Accuracy refers to the fraction of correctly classified instances to the total number of instances. Precision refers to the fraction of predicted relevant instances that are correctly classified as relevant. Recall refers to the fraction of the relevant instances that are correctly identified as relevant. F1-score is the weighted average of precision and recall. Another performance measure, Mean Average Precision (MAP) is used to evaluate the performance of ranking. It measures whether all relevant items are ranked highly. MAP is a preferred metric when not only the prediction or recommendation of relevant item but also their rank is important. Root Mean Square Error (RMSE) is another measure which has been used to find the difference between actual value and predicted value. Some researchers visualized the performance of the model using Area Under Curve (AUC), Area Under Precision Recall Curve (AUPRC), and Receiver Operating Characteristic (ROC) curve. In Table 2.3, we have listed the metrics used in different research works for evaluating their models.

2.5 Retweet for Information Diffusion

Prediction of retweet is an important topic of research because of its importance in the process of information diffusion. Now-a-days data from social networks is a great source of information. This information can be more useful when it can reach to appropriate users. Retweet has been used to determine trend and popularity of event (Gupta et al., 2012; Zhang P. et al., 2013; Hong et al., 2011). It was assumed that if an event gets relatively more tweets than retweets then the event might not last long on Twitter and in turns might not become popular. Gupta et al. (2012) checked the ratios of retweets at consecutive hours to capture the changes in popularity of the event over time. Zhang P. et al. (2013) used retweet to predict Twitter trend and Hong et al. (2011) explored retweets as a measure to find popular messages. Research showed that retweets are used vastly by the users to spread disaster related useful information during emergency (Starbird & Palen, 2010; Kogan et al., 2015). Kogan et al. (2015) explored the

retweeting pattern of geo-vulnerable users during hurricane Sandy in year 2012. After checking the retweeting activity of geographically vulnerable users during four different time frames (before, during, short-after, and long after the disaster), they found that the size (based on nodes and edges) of retweeting network during the disaster is bigger than the size before and after the disaster. They also determined the important nodes of each time-sliced retweet network using PageRank method and found that local government authorities and media are the most important nodes (most retweeted) in Geo-During network (formed by the retweets of geographically vulnerable users). Starbird and Palen (2010) also explored the use of retweet during two emergency situations - “Red River Flooding (USA), 2009" and "Oklahoma Fires (USA), 2009". This research also indicated that during emergency, tweets of local users, media and service organizations as well as tweets containing emergency related terms were retweeted more.

Contribution of retweet to engage remote individuals in 2011 Egyptian political uprising has been explored by Starbird and Palen (2012). During this event, protesters, journalist, media on the ground used to post movement-related information which were vastly retweeted by others to spread the information. This study showed that some tweets were not authored by people from Cairo but got high number of retweets and revolution-related metaphors were highly propagated in the Twitter. These findings clearly indicate the use of retweets by the users from Twitter in support of revolution. Sanjari and Khazraee (2014) explored information diffusion using Twitter during 2013 Iranian Presidential election. They showed that Iranian Twitter celebrities are most influential during election based on their retweet network. On the other hand, discussion about Iran in English was dominated by journalists and official media. Stieglitz and Dang-Xuan (2012) identified that political discussion took place in Twitter through retweet and direct message functionality and few highly active users are most influential whose tweets were retweeted vastly. In this study they found that leftists are the most influential users (got highest retweets). Since positive or negative sentimental tweets have high retweetability, tweets containing political sentiments are retweeted more and thus influence political decision.

Retweets are not just a method of information diffusion; they can be considered as a measure of trust between author and user. Trust is an important factor in social network to assess the credibility of information as well as to understand the flow of information in social network.

A user retweets an author's tweet when he has trust on that author. Adali et al. (2010) showed that user explicitly retweeting an author's tweets is a reliable measure of trust between two users.

Retweet is an excellent medium of information diffusion, especially during the time of emergency. However, because of its easy availability, during crisis time, along with important information some rumours can also be spreading through this mechanism. Abdullah et al. (2015) did a research to explore users' actions and decision-making behaviour on retweeting which helped them explain the reasons for spreading rumours at crisis time. According to the survey conducted by the authors, when users just retweet a message finding it important or marking it as favourite, there is greater chance to spread inaccurate information. However, when users search for further information regarding the tweet (the current situation); there is less chance to spread rumours at the time of disaster. Acar and Muraki (2011) also suggested that, use of official hashtags and provision to trace the originality of information would be effective solution to handle misinformation as well as might increase the reliability of information.

2.6 Discussions

A lot of works have been done on retweet prediction, but there is scope to further explore the influence of user behavior on retweet prediction. This research tries to identify users' behavior that has impact on their retweet decision. Researchers have successfully identified interest similarity (between user and tweet) as an informative feature to predict retweets; personality as good feature to predict retweeters, and emotion-sentiment as indicator to find tweet's retweetability. The objective of this research is to find the combined effect of topic, emotion-sentiment, and personality to predict retweets. There are still many unexplored latent factors that can represent users' behavior such as values, beliefs, and views on topics. In this study, we could not use them because we could not find any database or lexicon that defines the relation between words a user used in his tweets and his values, beliefs, etc.

Currently we chose to use only a few explicit content-based features such as hashtag, URL, and usernames. We did not include other explicit features such as presence-absence of question mark, length of tweet, language because we assume hashtags, URL, and usernames carry more

in-depth information regarding a user's interest than other explicit content-based features. Moreover, our purpose is to study implicit features. Explicit features have been studied a lot in the past, and we can easily include them in the model. Here, we just picked a few (hashtags, URL, user-names) representative ones.

To build retweet prediction model using machine learning methods, we choose XGBoost because it showed superior performance in many machine learning challenges, and we choose random forest also because of its superior performance in many previous works, and we want to compare their performances to see whether there is difference and how big the difference is.

Objective of this research is to explore users' behavior that influences their retweet decision. Past researchers used emotion-sentiment to find out the retweetability of tweets (Naveed et al., 2011; Pfitzner et al., 2012; Stieglitz & Dang-Xuan, 2012; Jenders et al., 2013; Kim & Yoo 2012). They explored the effects of emotion-sentiment from global perspective. In this research, we wanted to explore effects of emotion-sentiment from local perspective because we believe emotion and sentiment are two very unique concealed human attitudes. A user's emotion-sentiment reflected by their past behavior could provide valuable information regarding their future preference. Past researchers only considered following two-dimensional emotion-sentiment: positive and negative. In this research we have considered 10-dimensional emotion-sentiment assuming a detailed representation of emotion-sentiment would provide more precise information regarding human behavior. Lee et al. (2015) used personality of users to find retweeters from strangers. They assumed that personality of a user has good impact on their retweet activity. Personality served as a good feature to predict retweeters, and it is a subtle attribute of human nature which leads them to take different actions. So, we believe personality can be considered as a potential feature to predict users' retweet decision. In our past research, we already used personality as a feature to predict retweets (Firdaus et al. 2016). But in this work, we want to explore the combined effect of personality with other latent factors. We also want to explore the performance of matrix factorization retweet prediction method because it learns the latent relations between user and message in the process of predicting retweet. Jiang et al. (2015) applied message clustering technique and Wang et al. (2015) used social relationship between users to constrain the objective function of retweet prediction model. Message

clustering technique needs to regenerate clusters each time a new message is added to the dataset and social-relationship-based regularization needs to estimate social relation between every pair of users. Both of these procedures are not computationally feasible for fast growing huge social networks. Therefore, new terms are introduced which calculate the similarity between a user's messages to constrain the objective function of retweet prediction model.

Chapter 3

Retweet Prediction Model

This chapter provides the description of methods which are used to develop retweet prediction models for this research. The first part describes the concept of explicit and implicit features which are used to build our retweet prediction models. According to our current work, the retweet decision is mainly made based on the tweet itself, users' preferences, interests and (mental) states when reading the tweet. Features of our retweet prediction model are based on content of the tweets, which reflects user preference and inner feelings for the content of the tweet. We do not consider the social factors in this work, for example, how close the user is to the author, how active their interactions are in the past, how many times the tweet has been retweeted before the user reads the tweet, what kind of social behavior the user shows (e.g., tend to retweet tweets that have been retweeted by many friends). Then, in the second part we explain machine learning and matrix factorization retweet prediction models.

3.1 Features in Retweet Prediction Model

3.1.1 Explicit Features

Explicit features are the attributes of tweets which are directly expressed or mentioned in the tweets. For example, user-mention, hashtag, Universal Resource Locator (URL), punctuation mark, and emoticon. These factors have been used successfully by previous researchers (Peng et al., 2011; Suh et al., 2010; Xu & Yang, 2012; Naveed et al., 2011; Uysal & Croft, 2011; Kim & Yoo, 2012). For our research, we use user-mention, hashtags, and URL as explicit features because of their importance in understanding users' preference. User mentions are words starting

with "@" and used to mention usernames in tweets. User-mentions are used to communicate, address or notify another user about the tweet. It is assumed that, if a user communicates with another user in the past, he might communicate with that user in future as well. URL defines the address of a web page, which is used by the author of the tweet to redirect reader to that web page. It is considered as a tool for sharing or spreading information contained in that webpage. A user might share his preferred information by using URL in the tweet. If a user shares a URL in the past, he might share the same URL in future as well. Hashtags are words starting with "#" symbol. It is used to associate tweet with a certain topic. Hashtags allow user to track and follow information on a certain topic. It is believed that if a user follows a topic in the past, he might follow the same topic in future. Since past user-mentions, hashtags and URLs are good indicator of a user's future preference; we checked the presence of these factors in tweets to find their preferred tweets when building our retweet prediction models.

3.1.2 Implicit Features

Implicit features are the attributes of tweets which are not directly expressed or mentioned in the tweets. These are the factors in tweets that are not directly deducible but have great impact on users' retweet decision. Third-party method is required to mine these implicit attributes of tweets. Some of the implicit factors include topic of a tweet as well as emotion and personality reflected by a tweet. Past researchers have used topic and 2-dimensional sentiment for retweet prediction. In our research, we choose topic keywords, 10-dimensional emotion, and 35-dimensional personality as the implicit features for prediction model. We believe that they are the underlying attributes that have strong influence on a user's retweet decision. Description of these implicit factors is given below.

3.1.2.1 Topic

Topic is considered as the subject of the post it is referring to. Different methods have been used to define the topic of the tweet. Hashtags (Kanavos et al., 2014; Roberts et al., 2012) have been used as the corresponding topic of the tweet. For example, Kanavos et al. (2014) considered tweets with hashtag #MH370 as tweets on topic Malaysia Airlines Flight 370. Researchers

mainly used Latent Dirichlet Allocation (LDA) to extract topics. According to LDA method, topic assignment of the document is an iterative process which checks and updates the topic assignment of each word in every document based on the following two criteria: how frequent the word occurs across topics, and how frequent the topic occurs in the document. Finally, the most appropriate topics are chosen for the document. There are a few tools and packages implementing LDA model that can be used to find topics in a document, for example, Stanford Topic Modeling toolbox[12] and Mallet[13] (Machine Learning for Language Toolkit) topic modeling tool. When LDA is used to create topic-based user profile, a user's all past tweets are considered as a single document and LDA is used to find the topic distribution of that document (Xu & Yang, 2012). When LDA is used to identify topic of the tweet, the single tweet is considered as a document and LDA finds the distribution of topic in that document (Zhang J. et al., 2013). Since the original LDA model was developed for long document and might not work properly for short document like tweets, Zhao et al. (2011) proposed Twitter-LDA which is an extension of original LDA. We used Twitter-LDA to extract topics from the tweets. This extended version of LDA determines a single topic for a tweet. It is reasonable to assume that tweets are focusing on a single topic because of their short length. It is assumed that, there are T topics in Twitter where every topic is represented by a word distribution. There is also a word distribution for background model and topic distribution for every user. Since each tweet is generated by single topic and background model, in case of tweet generation process, a user first picks a topic based on topic distribution for user. Then words for the tweet are chosen one by one based on the selected topic or background model. This word selection process is directed by Bernoulli distribution.

3.1.2.2 Emotion and Sentiment

Emotion is considered as one of the most surprising and challenging factors of human psychology because users might show different emotions on a same topic. One user might think

[12]http://nlp.stanford.edu/software/tmt/tmt-0.4/

[13]http://mallet.cs.umass.edu/topics.php

positively about a topic whereas the other might think negatively. It is regarded as specific mental state towards an object or circumstance. The study of human emotion is necessary to understand user behavior and day to day activities. Users' social interaction and communication are influenced by emotion. For example, a sentence "Thank you dear!" is an expression of happiness and joy. Happiness might lead to initiate more interaction between two people. Though it is believed that human emotions are usually expressed through different facial expressions, emotions can also be expressed through a person's written content. Successful research has been done to find users' emotions and sentiment from their written documents (Mohammad & Turney, 2013; Bravo-Marquez et al., 2016; Roberts et al., 2012). We consider emotion as an effective subtle driving factor in user's retweet decision. We use a user's emotions and sentiments reflected by their tweets as factors for retweet prediction model. A number of theories have been proposed by the psychologists that classify human emotions. For this work, we are using well-accepted Plutchik's wheel of emotion theory (Plutchik, 2001) which classifies human emotions into eight basic categories: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Plutchik's wheel of emotion theory also mentioned some dyad emotions which are the mixture of two primary emotions such as hatred, submission, and disapproval. For example, emotion *hatred* is initiated by *disgust* and *anger*. For this study, we only consider eight basic emotions. We also use sentiment as a representative of a user's behavioral pattern for the purpose of retweet prediction. Past researchers (Pfitzner et al., 2012, Stieglitz & Dang-Xuan, 2012) considered positive or negative sentiment as emotion. But emotion and sentiment are two different mental states. Emotion is one's strong mental feeling and sentiment is their mental attitude triggered by emotion. For example, emotion *joy* triggers positive sentiment whereas *disgust* triggers negative sentiment. Sentiment is classified as either positive or negative. Positive sentiment refers to positive opinion or attitude whereas negative sentiment refers to the opposite.

For this research, we have decided to use 10 dimentional emotion-sentiment because we followed Plutchik's wheel of emotion theory which is well-accepted; and we also got previously published well-accepted lexicons developed based on Plutchik's wheel of emotion theory which give word-emotion association scores. To find the emotion and sentiment reflected by tweets, we used the following two word-emotion lexicons: NRC Word-Emotion Association Lexicon proposed by Mohammad and Turney (2013) and Expanded Version of NRC Word-Emotion

Association Lexicon proposed by Bravo-Marquez et al. (2016). Mohammad and Turney (2013) used Amazon's Mechanical Turk service to utilize the potency and intelligence of people to generate a large high-quality word-emotion association lexicon. This lexicon annotates 14182 distinct English words in eight basic Plutchik's wheel of emotion and two basic sentiment categories. NRC Word-Emotion Association Lexicon defines association of every word with each of eight emotions and two sentiments by association flag. Association flag value "0" denotes no association between word and the emotion/sentiment category whereas association flag value "1" indicates an association. A word can be associated with a single emotion or multiple emotions. For example, "asserting" is associated with only *trust* whereas "eager" is associated with *anticipation*, *joy*, *surprise*, and *trust*. Mohammad and Turney (2013) did pilot study on two different types of annotations: evokes, and association. In case of evokes, annotators were asked whether the word evokes a certain emotion. In case of association, annotators were asked whether the word is associated with a certain emotion. According to their study result, in most cases annotators agree on same word-emotion association whereas word evoking emotion is different for different annotators. So, the final word-emotion lexicon proposed by Mohammad and Turney defines association of every word with different emotion.

Bravo-Marquez et al. (2016) extended this lexicon for Twitter data. Since the general lexicon does not cover informal and misspelled words frequently used in Twitter, Bravo-Marquez et al. (2016) proposed a lexicon which is specific for language used in Twitter. For this expansion, they extracted different word-level features such as unigrams, Brown clusters, POS tags, and word2vec embeddings from tweets and used multi-label classification of word to categorize words in different emotions. They showed that the expanded lexicon performed better than the original NRC Word-Emotion Association Lexicon for finding the emotion categories reflected by tweets. This expanded lexicon gives the association score of each word with each of eight emotion and two sentiment categories. For example, word-emotion association scores of word *noooo*'s 10-dimensional emotion-sentiment*: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust* are listed in the following vector,

$$AssociationScore == \{0.7622, 0.\ 02158, 0.7997, 0.5482, 0.0055, 0.9817, 0.0143, 0.6339, 0.0516, 0.0065\}$$

For our study, we have decided to use both lexicons to gain the strength of both. Twitter based lexicon was mainly used to find the emotions reflected by tweets. During preliminary analysis with Twitter based lexicon, we found that for some tweets the scores for different emotion dimensions are very close to each other; this might give confusing result. To avoid this confusion, we decided to use word-emotion lexicon from Mohammad and Turney (2013) to amplify the effect of dominating emotions in a tweet. Details are explained in later sections.

3.1.2.3 Personality

Personality is a special trait of human being which forms a unique behavior pattern, feeling, and thought process for every individual. Personality is a subtle attribute of human character which has great impact on users' decisions and activities. We use personality as a feature for retweet prediction because personality can be considered as a representative behavior of a user which covers broad spectrum of users' exposed and latent nature. In the past few years, several successful works have been done to predict the personality of social network users based on their activities in the network (Golbeck, Robles, Edmondson, & Turner, 2011; Adali & Golbeck, 2012; Golbeck, Robles, & Turner, 2011; Quercia et al., 2011; Qiu et al., 2012; Li et al., 2014). The results from these research works have shown that personality is closely related to users' activities in social networks. As a measure of personality, we calculate Big Five and their thirty lower level facet scores for each user. Big Five is a widely accepted measure by the psychologist to model an individual's personality (Costa & McCrae, 1992). Big Five model defines an individual's personality in five different dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. High score in a dimension represents an individual's positive or strong nature in that trait whereas low score represents their opposite nature. Each of these dimensions defines a specific quality of human character. For example, *openness* refers to an individual's curious and creative nature whereas *neuroticism* refers to their temperament and mood. We also use thirty lower order facets of Big Five personality traits as described in Yarkoni (Yarkoni, 2010). Thirty lower order personality facets are: anxiety, anger, depression, self-consciousness, immoderation, vulnerability, friendliness, gregariousness, assertiveness, activity level, excitement-seeking, cheerfulness, imagination, artistic interests, emotionality, adventurousness, intellect, liberalism, trust, morality, altruism,

cooperation, modesty, sympathy, self-efficacy, orderliness, dutifulness, achievement striving, self-discipline, and cautiousness.

For this research, we have decided to use 35 dimentional personality vector to describe individual's personality reflected by tweets because we followed Big 5 personality theory. Big 5 is a well-accepted measure to model an individual's personality; and a highly accepted publication by Yorkani (2010) provided correlation scores between LIWC word-categoty and Big 5 and their 30 lower level facets. Therefore, to calculate 35 dimensional personality scores reflected by tweets and retweets, we use word-category and personality correlation scores proposed by Yarkoni (Yarkoni, 2010). Yarkoni used a large sample of blogs to find the association between personality and written language. This study presented the association between 35 personality traits and 66 Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) category. The psychologically meaningful LIWC categories used by Yarkoni (Yarkoni, 2010) are as follows: total pronouns, first person singular, first person plural, first person, second person, third person, negations, assent, articles, prepositions, numbers, affect, positive emotions, positive feelings, optimism, negative emotions, anxiety, anger, sadness, cognitive processes, causation, insight, discrepancy, inhibition, tentative, certainty, sensory processes, seeing, hearing, feeling, social process, communication, other references, friends, family, humans, time, past tense verbs, future tense verbs, space, up, down, inclusive, exclusive, motion, occupation, school, job/work, achievement, leisure, home, sports, TV/movies, music, money/finance, metaphysical, religion, death, physical states, body states, sexuality, eating/drinking, sleep, grooming, and swear words. For example, correlation score between LIWC category *music* and personality trait *extraversion* is 0.13 and correlation score between LIWC category *leisur*e and personality trait *openness* is -0.17.

3.2 Machine Learning Based Retweet Prediction Model

This section describes the system architecture of our retweet prediction model using machine learning algorithms. The description of strategies used to generate the model is then given.

3.2.1 System Architecture

The architecture of the proposed retweet prediction model is shown in Figure 3.1. It consists of mainly two components: storage component and application component. Storage component stores users' tweets and retweets. The application component accesses tweets and retweets from storage component and processes them for further tasks. Application component is further divided into several sub-components and their functionalities are explained below.

The data pre-processor prepares users' tweet and retweet data for further processing. The major pre-processing step is to tokenize every user's tweets and retweets. We do not include any stop-word removal or stemming step to use the data for our further profile generation step. During profile generation step, the personality score generation task uses the correlation coefficient values between different personality traits and LIWC categories as proposed by Yarkoni (Yarkoni, 2010). A few of the LIWC categories such as pronouns, first person singular, articles, preposition include stop-words, and thus, we do not remove the stop-words from the file. The LIWC program uses LIWC dictionary which includes word stems along with words. It finds the percentage of words or word stems in the data file which belong to a certain LIWC category. So, we do not stem any word beforehand. Another task of profile generation step is emotion and sentiment score generation. For this task, we use word-emotion association lexicon (Bravo-Marquez et al., 2016; Mohammad & Turney, 2013). To keep the words relevant for these lexicons, we do not remove stop-words or stem any words as well. Profile generation step also includes topic extraction task. To keep the data consistent with data used for personality, emotion, and sentiment generation tasks, we do not use any stop-word removal or stemming method during topic extraction task.

The profile generator generates a user's profile based on their past posts. It also generates profile for each tweet in the test set. In this work, each tweet in the test set is defined as target tweet. User profile is generated from their past posts and is defined by explicit and implicit features of tweets. Explicit features include user-mentions, URLs, and hashtags. Implicit features include topic, emotion, and personality. Target tweet profile is also defined by its explicit (user-mention, URL, hashtag) and implicit features (topic, emotion, personality). Six functions are implemented in this subcomponent. User-mention extractor extracts user-mention words (if any)

from a user's past posts. URL extractor extracts URL (if any) from a user's posts. Hashtag extractor gets hashtags (if any) from a user's posts. Topic extractor determines topic of each post. We use Twitter-LDA (Zhao et al., 2011) to extract topics of the posts. Each topic is defined by a set of keywords. Emotion calculator calculates the 8-dimensional emotion and 2-dimensional sentiment score of the tweet. It uses well-accepted word-emotion lexicons (Mohammad & Turney, 2013; Bravo-Marquez et al., 2016; Yarkoni, 2010) to find the emotion and sentiment score of each dimension reflected by the text in tweets. Personality score generator generates the 35-dimensional personality vector reflected by the posts. The difference between user profile and target tweet profile is, in case of user profile, this component processes all the past posts whereas for target tweet profile, it only considers the target tweet. Similarity calculator calculates similarity between user profile and target tweet profile. This component uses Jaccard and cosine similarity method to measure similarity between user and target tweet profile. The last component is retweet predictor. It uses the similarity scores generated from similarity calculator as features for retweet prediction. Finally, a classifier predicts the decision of retweet based on the features.

The retweet prediction includes three major tasks. The first task is to process users' past posts and generate the profile for every user. The second task is to process the target tweet and generate the tweet profile. And the third is to generate features and use a machine learning algorithm to predict whether the target tweet will be retweeted by the user. The workflow for the first task could be described as follows: (1) for every user, all of their past posts in two months (range from 60 to 2000) are accessed from storage, and then each of these posts is pre-processed; (2) user-mention, URL, and Hashtag are extracted from every past post of the user, (3) topic of a user's every post is determined, (4) 10-dimensional emotion-sentiment score vector is generated for all past posts, (5) 35-dimensional personality score vector is generated for all past posts. The workflow for the second task is as follows: (1) every target tweet is pre-processed; (2) user-mention, URL, and hashtag are extracted from the target tweet, (3) topic of the target tweet is determined, (4) 10-dimensional emotion-sentiment score reflected by the target tweet is calculated, (5) 35-dimensional personality score reflected by the target tweet is generated. In the workflow of the third task, features for retweet prediction model are generated using user profile and target tweet profile and those features are used by the retweet predictor to make retweet

decision. As a binary classification problem, for every user-target tweet pair, the retweet predictor classifies the target tweet as positive retweet (potential retweet by the user) or negative retweet (potential non-retweet by the user). Figure 3.2 shows the flowchart of the workflows of retweet prediction task.

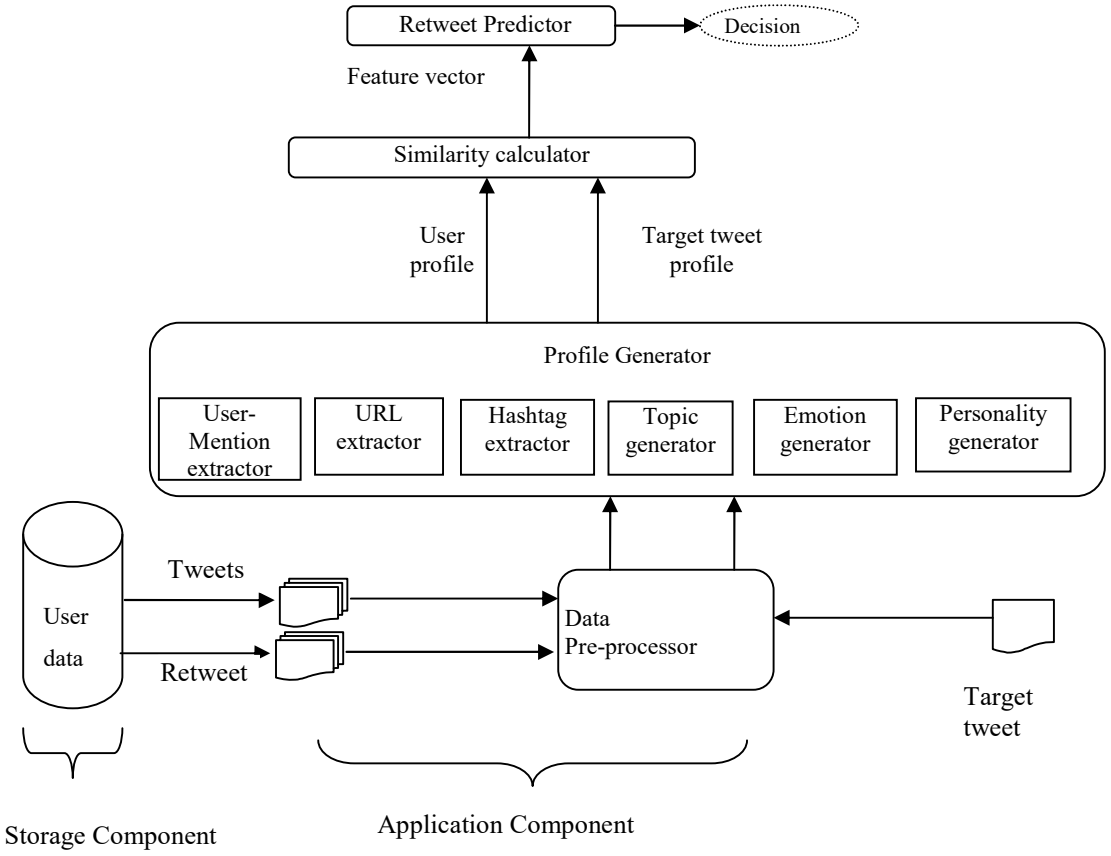
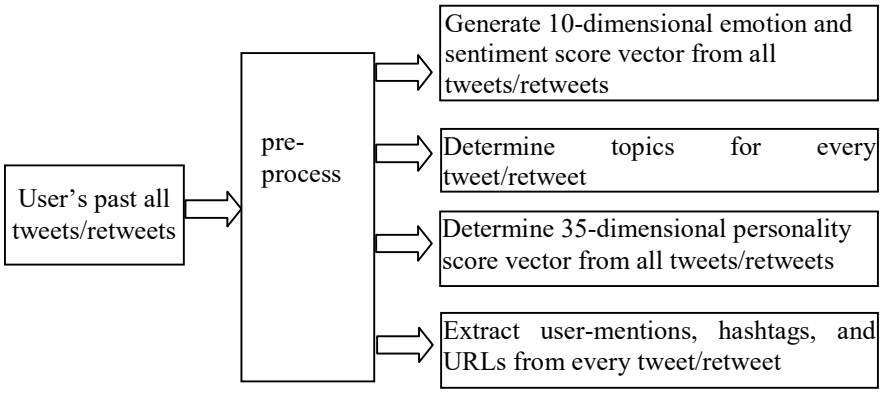
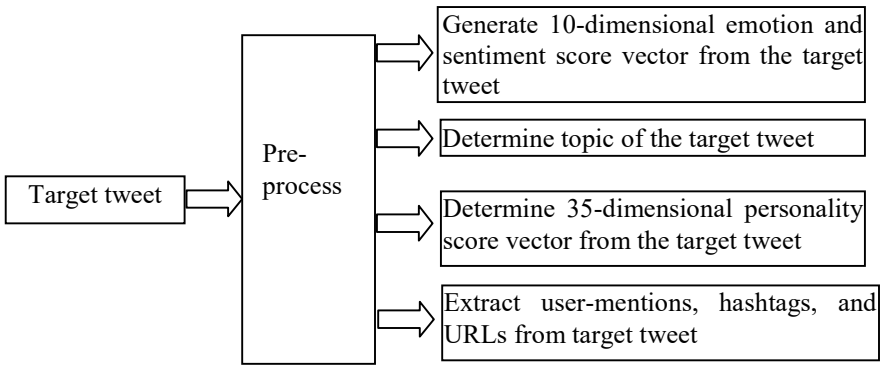


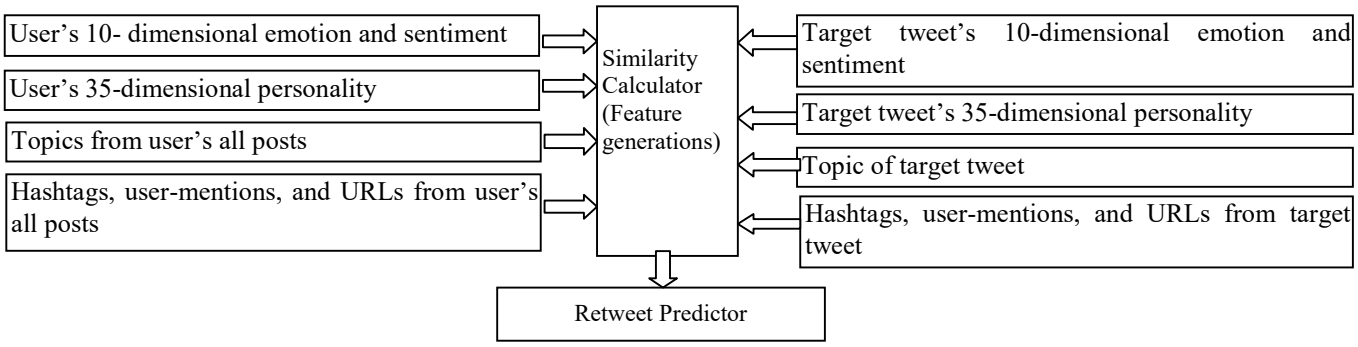
Figure 3.1: Architecture of retweet prediction model



(a)



(b)



(c)

Figure 3.2: Flowcharts showing the three major tasks of retweet prediction model:(a) Workflow of the 1st task that processes past tweets/retweets to create user profile; (b) Workflow of the 2nd task that processes every target tweet; (c) Workflow of the 3rd task that generates features and predicts the outcome.

3.2.2 Profile Generation

The strategies used by profile generator to process data in order to get desired scores are described in this section. User profile and target tweet profile are generated using the same strategies. The only difference is, in case of user profile, their all past posts are considered to create the profile whereas only the target tweet is considered to create target tweet profile. A user profile is represented by their previously used hashtags, user-mentions, URLs as well as topic keywords, emotion-sentiment, and personality reflected by all their past posts. Target tweet profile is represented by its hashtag (if any), user-mention (if any), URL (if any) as well as topic keywords, emotion-sentiment, and personality reflected by the target tweet. For this study we want to explore whether the retweet-only user profile performs better than tweet-only or tweet-plus-retweet user profile. So, we generate three types of user profiles. In case of tweet-only profile, a user's past tweets are used to create their profile. A user's past retweets are used to create retweet-only user profile. For tweet-plus-retweet profile, a user's both past tweets and retweets are used to create the profile.

3.2.2.1 Emotion and Sentiment Score Generation

Emotion and sentiment score reflected by a user's past posts or target tweet is represented by a 10-dimensional vector (8-dimension for emotion and 2-dimension for sentiment). In the rest of the thesis, we will simply refer to this as 10-dimensional emotion vector unless we specify differently. Suppose U is a set of n users, $U = \{u_1, u_2, u, \dots\dots, u_n\}$, For k^{th} user u_k, all his past posts are combined together and represented as tw_k, and the corresponding 10-dimensional emotion vector for tw_k is E_k, $E_k = \{e_{k1}, e_{k2}, e_{k3}, \dots\dots, e_{k10}\}$. For a user u_k, emotion scores reflected by their posts tw_k are calculated as follows:

i. Calculate the score of each emotion dimension j by adding the association score of each word w_m in tw_k, if it is found in Word-Emotion Association Lexicon (Bravo-Marquez et al., 2016) for that dimension.

$$ES_k = \{S_{k1}, S_{k2}, S_{k3}, \dots\dots, S_{k10}\} \quad (3.1)$$

$$S_{kj} = \sum_{m=1}^{N} AssScore_{mj} \quad (3.2)$$

where ES_k is the emotion association score vector for tw_k, S_{kj} is the score on emotion dimension j for tw_k, $AssScore_{ij}$ is the word-emotion association probability score between m^{th} word and j, and N is total number of words in tw_k.

ii. Find the number of words in tw_k which belong to each of the emotion dimensions j using NRC word-emotion lexicon (Mohammad & Turney, 2013), NW_k=$\{n_{k1}, n_{k2}, n_{k3},.., n_{k1}\}$, where n_{kj}= number of words in tw_k belonging to emotion dimension j.

iii. The final emotion score for tw_k for emotion dimension j is calculated using equation (3.3).

$$e_{kj} = \begin{cases} S_{kj} * n_{kj}, & if\, n_{kj} \neq 0 \\ S_{kj}, & if\, n_{kj} = 0 \end{cases} \quad (3.3)$$

Finally, the scores for every emotion dimension were normalized. As explained earlier, we used the second lexicon to amplify the effect of dominating emotions. For example, check the emotion-sentiment reflected by the following tweet "Thanks! Happy mother's day". Based on word-emotion lexicon proposed by Bravo-Marquez *et al.* (2016), the scores for 10-dimensional emotion for the tweet (using equation 3.1 and 3.2) are: anger: 0.0158, anticipation: 0.5136, disgust: 0.04683, fear: 0.42738, joy: 0.89035, sad: 0.08141, surprise: 0.09465, trust: 0.66133, negative: 0.06914, positive: 1.32493. Using only this word-emotion lexicon, the prominent emotion (score greater than median) reflected by the text are anticipation, fear, joy, and trust. Intuitively we can say that the given text reflects joy, anticipation, and trust; but not fear. Now we use the word-emotion lexicon proposed by Mohammad and Turney (2013) to find the number of word association with different emotion, we found that 2 words are associated with each of the following emotion: anticipation, joy, trust. No words are actually associated with fear. So, when we multiply the emotional dimensional score with number of associated words in that dimension (using equation 3.3), the gap between the scores for anticipation, joy, and trust and scores for other dimensions becomes bigger. As per the final scores, the prominent emotions

reflected by the text are anticipation, joy, and trust, which also complys with our natural intuition about the text.

3.2.2.2 Personality Score Generation

For every user, we calculate their 35-dimensional personality vector which consists of their Big Five and thirty lower level personality scores based on their past posts. Here we do not involve any human subject directly to measure their personality scores. We follow a well-accepted process proposed by Lee et al. (2015) to calculate personality scores of social network users. Personality of a user is calculated based on their tweets and retweets. The first step in this process is to use LIWC program to find the percentage of words in a user's file which belong to some psychologically meaningful categories. For the purpose of personality calculation, we segment the hashtags used in tweets to find the meaningful category of the hashtags. For example, hashtag #ProudCanadian is segmented as Proud Canadian, so that these two words can be considered when calculating percentage of words in LIWC category. In case of calculating emotion reflected by the tweet, we do not segment the hashtags because the used word-emotion association lexicon (Bravo-Marquez et al., 2016) provided association scores between different hashtags and emotion dimension. Personality score of each of the Big Five traits and their thirty sub-dimensions was calculated using correlation coefficient value between LIWC categories and the respective Big Five (or their thirty sub-dimension) trait. LIWC program (using 2001 dictionary) specifies more than 70 different categories. But we kept only 66 categories which have correlation with personality traits as specified by Yarkoni (Yarkoni, 2010). Then we calculate 35-dimensional (Big 5 dimension + 30 sub-dimensions) personality vector of each user based on his past posts. Suppose we use U to represent the set of n users, $U = \{u_1, u_2, u_3, ..., u_n\}$. For k^{th} user u_k, his 35-dimensional personality vector based on posts is,

$$Per_k = \{p_{1k}, p_{2k}, p_{3k}, \dots\dots, p_{35k}\}$$

where p_{ik} is the i^{th} personality trait for user u_k. For a user u_k, score for i^{th} personality trait can be calculated using the following equation.

$$p_{ik} = \sum_{j=1}^{66} N_{jk}\, C_{ij} \quad (3.4)$$

In (3.4),

N_{jk}= percentage of words in u_k's posts which belong to j^{th} LIWC category

C_{ij}= correlation value between j^{th} LIWC category and personality trait p_{ik}

Finally, scores for personality traits are normalized to be used in the prediction model. In equation 3.4, we use the correlation value between personality trait and LIWC category which was given by Yarkoni (Yarkoni, 2010).

3.2.2.3 Topic Generation

Basic LDA model was developed for longer documents and it might not work well for short text like tweets. Also, because of the length of a tweet, it is usually focused on one single topic. Zhao et al. (2011) extended the basic LDA model and proposed Twitter-LDA which determines a single topic for a tweet. In our work, we used Twitter LDA to determine topic of each tweet. Assuming there are *T* topics in Twitter where each topic is represented by a word distribution ϕ^t. Let $\phi^B \sim Dir(\beta)$ refers to the word distribution for background model and $\theta^u \sim Dir(\beta)$refers to the topic distribution for user *u*. In case of writing tweets, a user first picks a topic based on θ^u, then he chooses words one by one based on the selected topic or background model. This word selection step is governed by Bernoulli distribution denoted by $\pi \sim Dir(\gamma)$. Assuming each tweet is generated by a single topic and a background model, Zhao *et al.* (2011) described the process of tweet generation as follows:

1. Choose $\phi^B \sim Dir(\beta)$, $\pi \sim Dir(\gamma)$
2. For each topic $t = \{1, 2, 3, \ldots\ldots, T\}$
   Choose $\phi^t \sim Dir(\beta)$
3. For each user $u \in U$
   i. choose $\theta^u \sim Dir(\alpha)$
   ii. for each tweet $s = 1, \ldots, N_u$
   - Choose $z_{u,s} \sim Multi(\theta^u)$

- for each word $n = 1, \ldots, N_{u,s}$
  - $y_{u,s,n} \sim Multi(\pi)$
  - choose $w_{u,s,n} \sim Multi(\phi^{B})$ if $y_{u,s,n} = 0$ and $w_{u,s,n} \sim Multi(\varphi^{z_{u,s}})$ if $y_{u,s,n} = 1$

Here *Dir* indicates Dirichlet and *Multi* indicates Multinomial distribution. Along with topic for every tweet, Twitter-LDA also provides keywords representing every topic. These keywords provide in-depth information regarding topic of the tweet in broad spectrum. So, to calculate features for retweet prediction model, we use these keywords as a representation of the topic.

3.2.2.4 Feature Vector Generation

To predict whether a target tweet will be retweeted by a user, a 6-dimensional feature vector is prepared for every (target tweet, user) pair as the input to the prediction model. Jaccard and Cosine similarity methods are used to generate these features. Jaccard similarity measure is used for binary features when the only known information is whether a feature (such as a URL, a hashtag, a topic keyword) is in or not in a profile. Cosine similarity measure is used for non-binary features, for example, a vector of real-valued emotion scores. Among all the features, cosine similarity is used for emotion and personality based features, and Jaccard similarity is used for all other features. Every user-tweet pair is represented by a 6-dimensional feature vector $\{f_{um}, f_{url}, f_{ht}, f_{tp}, f_{em}, f_{per}\}$, which is described in Table 3.1. Methods of Jaccard and Cosine similarity measures are described below.

Jaccard similarity measure:

Jaccard similarity measures the similarity between user profile defined by a set of terms extracted from past posts and target tweet profile defined by a set of terms extracted from the target tweet. For example, Jaccard similarity between user profile consisting of previously used hashtags and target tweet profile consisting of hashtags is calculated using equation (3.5).

$$JaccardSimilarity = \frac{|A \cap B|}{|A \cup B|} \tag{3.5}$$

where A= set of hashtags from user profile

B= set of hashtags from target tweet profile

For example,

A={#Canada, #Trump, #Trudo, #iPhoneX, #Halloween, #MeToo, #trumpcare, #fun}

B={#Halloween, #spooky}

$$JaccardSimilarity = \frac{|\{\#Halloween\}|}{|\{\#Canada,\#Trump,\#Trudo,\#iPhoneX,\#Halloween,\#MeToo,\#trumpcare,\#fun,\#spooky\}|}$$

$$= \frac{1}{9} = 0.11$$

Cosine similarity measure:

Cosine similarity measures the similarity between user profile and target tweet profile which are defined by the vectors. For example, cosine similarity between 10-dimensional emotion vectors from the user profile and target tweet profile is calculated using equation (3.6).

$$CosineSimilarity = \frac{A \cdot B}{\|A\|\|B\|} \quad (3.6)$$

where, A = 10-dimensional emotion vector from the user profile

B = 10-dimensional emotion vector from the target tweet profile

For example,

A= {0.08273,0.07439,0.10819,0.34773,0.06332,0.02244,0.08924,0.21197,0.36124,0.63876}

B= {0.06141,0.12359,0.08048,0.22205,0.19390,0.07550,0.04190,0.20112,0.48950,0.51049}

$$CosineSimilarity = \frac{A \cdot B}{\|A\|\|B\|}$$

$$= \frac{\sum_{i=1}^{10} A_i B_i}{\sqrt{\sum_{i=1}^{10} A_i^2}\sqrt{\sum_{i=1}^{10} B_i^2}} = 0.9483$$

Table 3.1: Description of feature vector

Feature	**Description**
f_{um}	Jaccard similarity between user profile consisting of previously used user-mentions and user-mentions from target tweet
f_{url}	Jaccard similarity between user profile consisting of previously used URLs and URLs from target tweet
f_{ht}	Jaccard similarity between user profile consisting of previously used hashtags and hashtags from target tweet
f_{tp}	Jaccard similarity between user profile consisting of previously used topic keywords and keywords of target tweet's topic
f_{em}	Cosine similarity between user profile consisting of 10-dimensional emotion-sentiment scores and 10-dimensional emotion-sentiment scores of the target tweet
f_{per}	Cosine similarity between user profile consisting of 35-dimensional personality scores and 35-dimensional personality scores of the target tweet

3.2.3 Machine Learning Methods

Machine learning is an area of computer science where statistical methods are used to provide learning experience to computer systems. A well-established definition for machine learning algorithm given by Tom M. Mitchell (Mitchell, 1997) is stated below:

"*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.*"

The basic idea is to learn from past data effectively such that the future can be predicted by the method. Machine learning methods have become an essential part of computer science because of its effectiveness in handling highly complex real word problems. Machine learning methods are used for different prediction tasks such as event prediction, customer interest prediction, stock market prediction, disease prediction, weather prediction, retweet prediction etc.

Machine learning tasks are mainly divided in three categories: supervised leaning, unsupervised learning, and reinforcement learning (Sugiyama 2015). In case of supervised learning, the program is trained with a predefined (input with known target label) set of examples known as training data such that it learns rules to predict output for unknown input in future. In case of unsupervised leaning, the program is trained with examples where the target labels are not known. Unsupervised leaning finds the hidden pattern in the data. In case of reinforcement learning, the program learns from actions and corresponding results.

For our research, we use supervised learning method because our program is trained on a set of retweets where the target labels (retweets or non-retweets) are known. Target tweets are labeled as retweets when a user sees a tweet (posted by his followee) in their timeline and retweets it whereas target tweets are labeled as non-retweets when a user sees a tweet (posted by his followee) in their timeline but does not retweet it.

A large variety of machine learning methods are available to solve prediction problem. Some of the machine learning methods that can be used for prediction task are as follows:

Bayesian Statistics, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network. For this research, we used XGBoost (Extreme Gradient Boosting) and random forest algorithms to develop retweet prediction models. We used XGBoost because of its superior performance in different machine learning competitions such as Kaggle and KDDCup2015 (Chen & Guestrin, 2016). We used random forest algorithm for its excellent performance in various prediction tasks (Narayanan et al., 2011; Sumner et al., 2012; Arora et al., 2014; Fire et al., 2013; Can et al., 2013). We used these two algorithms to check whether the performance of the prediction model is similar or different when using different machine learning algorithms. Developing retweet prediction models can be considered as classification problem where the model classifies a target tweet as retweet or non-retweet.

3.2.3.1 XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable machine learning algorithm, proposed by Tianqi Chen, implements gradient boosting decision tree algorithm to solve machine learning problems. This scalable machine learning model is capable to handle sparse data and handle instance weights for approximate tree learning (Chen & Guestrin, 2016). XGBoost has showed its superiority for modeling classification and regression problems in different Kaggle competitions. XGBoost gave winning performance for several problems such as store sale prediction, web text classification, customer behavior prediction, product categorization, and malware classification (Chen & Guestrin, 2016). XGBoost is an ensemble classifier which builds models sequentially such that each model tries to reduce the errors of the past model. The process continues until there is no more improvement in performance. The basic boosting process consists of the following steps:

i. Create initial model to predict target variable. Find the error associated with the model
ii. Create a new model which fits the errors from the past step
iii. Add the new model to the initial one to create a boosted version of initial model. The error from this boosted model is lower than that of initial model.
iv. Continue steps (i) - (iii) until no more improvements in the minimization of error.

XGBoost uses gradient descent technique to minimize the error when adding new predictor, hence known as gradient boosting method. The inventors of XGBoost updated the traditional tree boosting method by adding a new regularization term in the objective function. This additional regularization term helps avoid overfitting by smoothing the final learnt weights (Chen & Guestrin, 2016).

3.2.3.2 Random Forest Algorithm

Random forest is a supervised classification algorithm which generates many decision trees and final class for a test sample is predicted based on the majority vote (Breiman, 2001; Gray et al., 2013). The training set of each tree is selected by bootstrap sampling. So, a random *N* samples are selected from the whole dataset as the training set for a tree in the forest. Approximately one third of the dataset is left behind which is known as out-of-bag data for the tree and can be used for internal test predictions. The generalization error of the random forest can be determined by aggregating the predictions of each of the trees in the forest. The basic steps of random forest algorithm are described below (Gray et al., 2013):

i. At each node in a tree, randomly select *k* features from the *N* available features where k << M.
ii. Partition the node using best possible binary split. Gini impurity is used to decide the best split.
iii. Repeat steps i and ii until the tree reaches to the maximum depth, i.e., tree having target node as leaf.
iv. Repeat steps i to iii for *n* number of trees to generate random forest.
v. For test example, do the prediction using rules from each tree in the forest.
vi. Calculate the vote for every predicted target and the target with highest vote is considered the final prediction.

3.3 Matrix Factorization Based Retweet Prediction Model

This section describes the basics of matrix factorization technique which is followed by the description of the proposed matrix factorization retweet prediction model.

3.3.1 Matrix Factorization Basic

Matrix factorization is a popular machine learning technique mostly used to predict user's preference based on their ratings for items and recommend items according to the predicted preference. The basic idea of matrix factorization is to factorize a matrix into two lower rank matrices such that multiplying them will reconstruct the original one. Different types of matrix factorization techniques are used in recommendation/prediction research. Most popular ones are Singular Value Decomposition (SVD), Regularized Singular Value Decomposition, Non-negative Matrix Factorization, Principal Component Analysis, and Factor Analysis.

The factorization process involves learning of latent features underlying the interactions between users and items. Latent features are hidden, non-human-extractable features which have the ability to define a user's preference or rating for an item. Learning latent features should be helpful to find a user's decision for a certain tweet based on the features associated with user and item. Learning the latent relation between user and item has made matrix factorization a popular machine learning approach for retweet prediction research (Wang et al., 2015; Jiang et al., 2015).

In basic matrix factorization technique, users' preferences for items are modeled by user-item rating matrix R which is defined by $R = R^{NxM}$ where N is the number of users and M is the number of items. R_{ij} = user i's rating for item j if user i rated item j; 0 otherwise. The rating matrix R is factorized into two low dimensional matrices user-factor matrix U and item-factor matrix V, $R = U$ x V where U is defined by $U = U^{NxK}$; V is defined by $V = V^{KxM}$; and K is the number of latent features. Each row of U defines the strength of association between user and latent features whereas each column of V defines item's association with latent features. The dot product $\mathrm{U_i V_j}$ defines user i's interest for item j where $\mathrm{V_j}$ represents $[\mathrm{V_{1j}}, \mathrm{V_{2j}}, \mathrm{V_{3j}}, \ldots, \mathrm{V_{kj}}]^{\mathrm{T}}$. So, user i's rating for item j can be predicted by equation (3.7).

$$\hat{R}_{ij} = \mathrm{U_i V_j} \tag{3.7}$$

The objective is to decrease the difference between actual rating R_{ij} and predicted rating $\hat{R}_{ij}$ known as prediction error. The main process includes learning of latent features by minimizing the regularized squared error on the known ratings defined by equation (3.8).

$$min_{U,V}\mathcal{J}(R,U,V) = \sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij}^2 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\,(\mathrm{R_{ij}} - \mathrm{U_i V_j})^2 + \frac{\beta}{2}\|U\|_F^2 + \frac{\gamma}{2}\|V\|_F^2 \qquad (3.8)$$

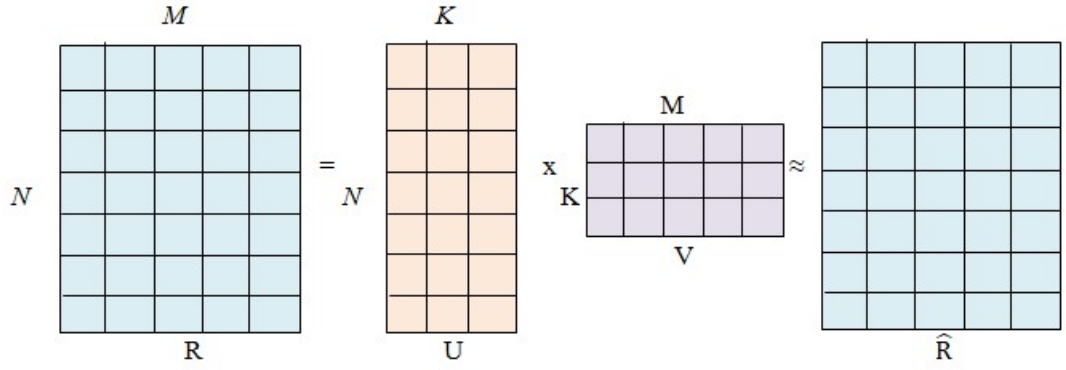


Figure 3.3: In matrix factorization, rating matrix $R = R^{NxM}$ is factorized to lower rank matrices $U = U^{NxK}$ and $V = V^{KxM}$ which are multiplied to reconstruct approximate of R.

In equation (3.8), $\|.\|_F$ denotes Frobenius 2-Norm. Two regularization parameters β and γ control the extent of regularization to avoid overfitting the system on the observed data. $I_{ij}\epsilon\{0,1\}$ where $I_{ij=}1$ if relation was observed between i and j; and I_{ij} =0 vice versa. This process uses stochastic gradient descent technique to obtain the local optimal solution. For every training case, the system measures the prediction error and update U and V in the opposite direction of the gradient to minimize the error. The gradient of the objective function defined by equation (3.8), with respect to $\mathrm{U_i}$ and $\mathrm{V_j}$ are described by equations (3.9) and (3.10).

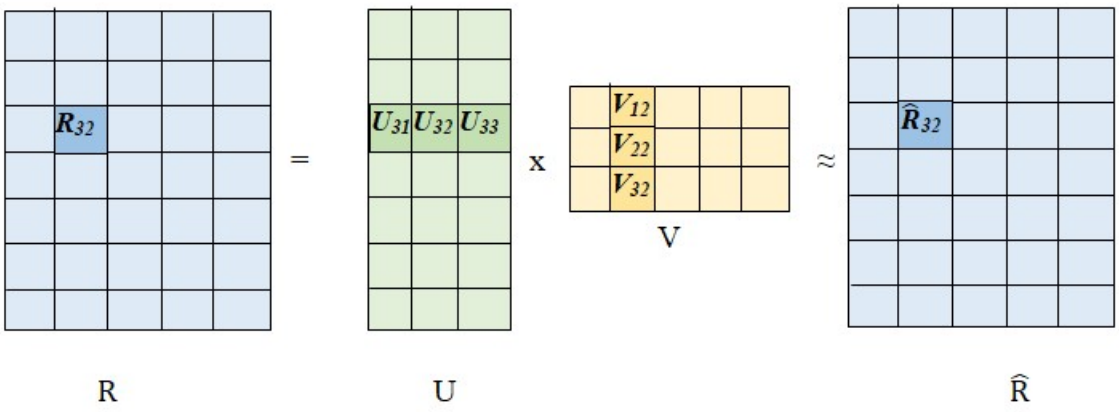
$$\frac{\partial \mathcal{J}}{\partial U_i} = \sum_{j=1}^{M} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})\left(-\mathrm{V_j}\right) + \beta\,\mathrm{U_i} \qquad (3.9)$$

$$\frac{\partial \mathcal{J}}{\partial V_j} = \sum_{i=1}^{N} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})\,(-\mathrm{U_i}) + \gamma\,\mathrm{V_j} \qquad (3.10)$$

In the next step, U_i and V_j are updated using the following equations where η is a small constant defines the learning rate while approaching to local optimum.

$$U_i = U_i - \eta \left(\frac{\partial J}{\partial U_i}\right) \tag{3.11}$$

$$V_j = V_j - \eta \left(\frac{\partial J}{\partial V_j}\right) \tag{3.12}$$



$$\hat{R}_{32} = (U_{31}, U_{32}, U_{33}) \,.\, (V_{12}, V_{22}, V_{32}) = U_{31} * V_{12} + U_{32} * V_{22} + U_{33} * V_{32}$$

Figure 3.4: Matrix factorization method to predict rating of 2nd item for 3rd user ($\hat{R}_{32}$)

3.3.2 Learning Algorithms for Matrix Factorization

Stochastic Gradient Descent (SGD) and Alternating Least Square (ALS) are two learning algorithms used in the optimization process of matrix factorization.

Stochastic Gradient Descent (SGD):

In this iterative optimization technique, for every training case, the algorithm computes the predicted rating, and then calculates the prediction error using equation (3.13) (Koren et al., 2009).

$$e_{ij} = R_{ij} - U_i V_j \tag{3.13}$$

Then it updates U_i and V_j by a value proportional to learning rate η using equations (3.14) and (3.15) (Koren et al., 2009).

$$U_i \leftarrow U_i + \eta\,(e_{ij} V_j - \beta\, U_i) \tag{3.14}$$

$$V_j \leftarrow V_j + \eta\,(e_{ij} U_i - \beta\, V_j) \tag{3.15}$$

Alternating Least Square (ALS):

Alternating Least Square is a two-step iterative optimization technique which can be useful when the process needs parallelization. At every iteration, this technique alternatively fixes U and V to solve the other (Koren et al., 2009). This two-step alternating process continues until the system reaches convergence. This method provides parallelization because computation of each U_i and V_j is not dependent on other user factors U_m where $m \neq i$ and item factors V_n where $n \neq j$.

3.3.3 Proposed Matrix Factorization Based Model

Two retweet prediction models have been developed using matrix factorization algorithm. In these models two new regularization terms are introduced based on message similarity to regularize the data for matrix factorization objective function. Previous research showed that users retweet messages which reflect their interest (Wang et al., 2015; Macskassy & Michelson, 2011). We assumed that messages retweeted by a user share similar characteristics in terms of content, whereas the non-retweeted messages by a user are dissimilar from the retweeted messages. Since a user's past activities (retweets and non-retweets) carry important information

about their future decision, new regularizers are individual user based. We also believe that if two messages are similar or dissimilar in the observed space, then this quotient will be consistent in the latent space. Based on this assumption, we have designed two individual user-based message similarity measures for regularization terms, which consider the similarity between retweets and non-retweets of a user to regularize the objective function for the purpose of getting local optimum solution. Jiang et al. (2015) also used message similarity based regularizer. To constrain objective function, they calculated the similarity between every pair of messages in a cluster whereas our proposed approach calculates similarity between a pair of messages from a same user. The algorithm and architecture of proposed matrix factorization retweet prediction models are given in Figure 3.5 and 3.6 respectively.

For retweet prediction, users' retweet behavior is modeled by user-message retweeting matrix R which is defined by $R = R^{NxM}$ where N is the number of users and M is the number of messages. $R_{ij} = 2$ if user i retweeted message j, $R_{ij} = 1$ if user i did not retweet message j (message j was in user i's timeline but they did not retweet it), and $R_{ij} = 0$ if there is no interaction between user and message. The retweet matrix R is factorized into two low dimensional matrices U and V. Each row of U defines the strength of relation between user and latent features whereas each column of V defines tweet's association with latent features. This work used Stochastic Gradient Descent (SGD) as learning algorithm to minimize the prediction error during the process of learning latent features. The objective of proposed matrix factorization retweet prediction model is to predict retweet relation between user and tweet.

3.3.3.1 Approach 1

The basic matrix factorization model is revised by adding a new regularization term using cosine similarity to measure the message similarity. We believe that, for a user, similarity between two messages in the latent space is consistent to their similarity in the observed space. The purpose of this regularization term is to reduce the difference between the similarity of a user's messages in the observed space and in the latent space. The main process includes learning the latent features by minimizing the regularized squared error. The basic regularized objective function of matrix factorization (equation 3.8) is revised by adding new regularization

term which is defined by equation (3.16). Algorithm and architecture of matrix factorization model are given in Figure 3.5 and Figure 3.6 respectively.

$$min_{U,V}\mathcal{J}(R,U,V) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\,(\mathrm{R_{ij}} - \mathrm{U_i V_j})^2 + \frac{\alpha_1}{2}\sum_{i=1}^{N}\sum_{j\in M_i}\sum_{k\in M_i}\,(Sim_{jk} - Cos(V_j,V_k))^2 + \frac{\beta}{2}\|U\|_F^2 + \frac{\gamma}{2}\|V\|_F^2 \quad (3.16)$$

Here, N is the number of users, M_i is the set of messages (positive/negative) related to i^{th} user, Sim_{jk} is the similarity between two messages j and k for user i in the observed space, and $Cos(V_j, V_k)$ is the cosine similarity between two messages j and k for user i in the latent space. $I_{ij} = 1$ if there is a relation between user i and message j; otherwise I_{ij}=0. Similarity calculation between two messages in observed space is described in Section 3.3.3.3.

The gradient of the objective function defined by equation (3.16), with respect to $\mathrm{U_i}$ and $\mathrm{V_j}$ are described by equations (3.17) and (3.18).

$$\frac{\partial \mathcal{J}}{\partial U_i} = \sum_{j=1}^{M} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})(-\mathrm{V_j}) + \beta\,\mathrm{U_i} \quad (3.17)$$

$$\frac{\partial \mathcal{J}}{\partial V_j} = \sum_{i=1}^{N} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})\,(-\mathrm{U_i}) + \gamma\,\mathrm{V_j} + \alpha_1 \sum_{k\in M_i}(Sim_{jk} - Cos(V_j,V_k)) \quad (3.18)$$

Update $\mathrm{U_i}$ and $\mathrm{V_j}$ using the following equations.

$$U_i = U_i - \eta\left(\frac{\partial \mathcal{J}}{\partial U_i}\right) \quad (3.19)$$

$$V_j = V_j - \eta\left(\frac{\partial \mathcal{J}}{\partial V_j}\right) \quad (3.20)$$

Input: Retweet matrix $R = R^{NxM}$; similarity between messages; maximum_step; Threshold; parameters: $\alpha_1, \beta, and\ \gamma$; Number of latent features K

Output: Predicted retweet matrix $\widehat{R}$

- Initialize user-factor matrix U=NxK and item-factor matrix V=KxM with random values.
- Repeat until step <= maximum_step:
    - for each $i \in \{1,2,3,\ldots..N\}$:
        for each $j \in \{1,2,3,\ldots.., M\}$:
            if $R_{ij} > 0$:
                $\hat{R}_{ij} = U_i V_j$
                calculate prediction error for element R_{ij}, $e_{ij} = R_{ij} - \hat{R}_{ij}$
                compute gradient of the objective function using equation (3.17) and (3.18)
                Update U_i and V_j using equations (3.19) and (3.20)
    - calculate regularized total error as of equation (3.16)
      if regularized total error < Threshold:
          break
- Return $\widehat{R} = U\ V$

Figure 3.5: Algorithm for proposed matrix factorization with approach 1

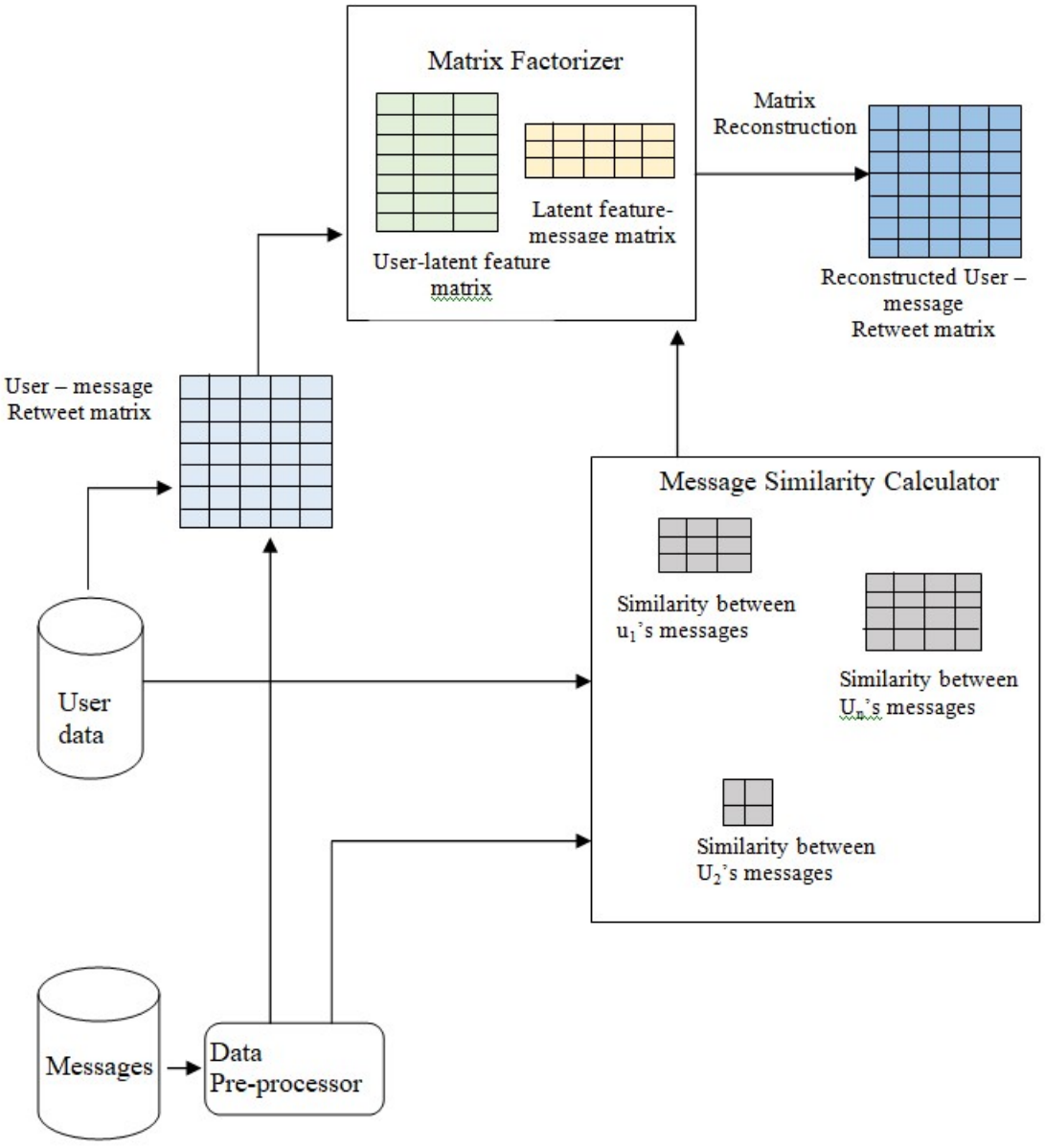


Figure 3.6: Architecture of matrix factorization retweet prediction model

3.3.3.2 Approach 2

Another regularizer based on individual user's message similarity is proposed. We assume that, if a user's two messages are similar in the observed space then the distance between their feature vectors would be smaller in the latent space. On the other hand, a smaller similarity score means larger distance between the feature vectors representing the messages. The purpose is to minimize the distance between the feature vectors which represent similar messages. The revised objective function with this new regularization term is described in equation (3.21).

$$min_{U,V}\mathcal{J}(R,U,V) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\,(\mathrm{R_{ij}} - \mathrm{U_i V_j})^2 + \frac{\alpha_2}{2}\sum_{i=1}^{N}\sum_{j\in M_i}\sum_{k\in M_i} Sim_{jk}\left\|V_j - V_k\right\|_F^2 + \frac{\beta}{2}\|U\|_F^2 + \frac{\gamma}{2}\|V\|_F^2 \quad (3.21)$$

Here, N is the number of users, M_i is the set of messages (retweets/non-retweets) related to i^{th} user, Sim_{jk} is the similarity between two messages j and k related to user i in the observed space. $I_{ij} = 1$ if there is a relation between user i and message j; otherwise I_{ij}=0. The regularizer used here is similar to the regularizer used by Jiang et al. (2015), but they calculated similarity of messages in a cluster whereas we used similarity between a user's messages.

The gradient of the objective function defined by equation (3.21), with respect to $\mathrm{U_i}$ and $\mathrm{V_j}$ are described by equations (3.22) and (3.23).

$$\frac{\partial \mathcal{J}}{\partial U_i} = \sum_{j=1}^{M} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})\left(-\mathrm{V_j}\right) + \beta\,\mathrm{U_i} \quad (3.22)$$

$$\frac{\partial \mathcal{J}}{\partial V_j} = \sum_{i=1}^{N} I_{ij}(\mathrm{R_{ij}} - \mathrm{U_i V_j})\,(-\mathrm{U_i}) + \gamma\,\mathrm{V_j} + \alpha_2 \sum_{k\in M_i} Sim_{jk}(V_j - V_k) \quad (3.23)$$

3.3.3.3 Calculation of similarity between two messages

Every message is represented by both explicit and implicit features. User-mentions, URLs, and hashtags are considered as explicit features of the message. Topic, emotion, and personality reflected by the message are considered as implicit features. Similarity between message m_i and m_j are calculated using equation (3.24).

$$Sim(m_i, m_j) = \delta * (Sim_{explicit}(m_i, m_j)) + (1 - \delta) * (Sim_{implicit}(m_i, m_j)) \quad (3.24)$$

Here, δ is a constant used to control the contribution of these features in the calculation of final similarity. The selection of δ is done after checking the performance of the model with different values. For explicit content similarity $Sim_{explicit}(m_i, m_j)$, Jaccard Similarity between the explicit features of the messages are calculated separately and then the scores are added. For implicit content similarity $Sim_{implicit}(m_i, m_j)$, cosine Similarity between the implicit features of the messages are calculated separately and then scores are added. The added explicit and implicit content similarity scores are divided by three to normalize the scores. Calculation of normalized $Sim_{explicit}(m_i, m_j)$ and $Sim_{implicit}(m_i, m_j)$ is described by the following equations.

$$Sim_{explicit}(m_i, m_j) = \frac{Jac_{um}(m_i, m_j) + Jac_{url}(m_i, m_j) + Ja_{ht}(m_i, m_j)}{3} \quad (3.25)$$

Here, $Jac_{um}(m_i, m_j)$ = Jaccard similarity between user-mentions of m_i and m_j

$Jac_{url}(m_i, m_j)$ = Jaccard similarity between urls of m_i and m_j

$Jac_{ht}(m_i, m_j)$ = Jaccard similarity between hahstags of m_i and m_j

$$Sim_{implicit}(m_i, m_j) = \frac{Jac_{tpKW}(m_i, m_j) + Cos_{em}(m_i, m_j) + Cos_{prsn}(m_i, m_j)}{3} \quad (3.26)$$

Here, $Jac_{tpKW}(m_i, m_j)$ = Jaccard similarity between the keywords of the topics of m_i and m_j.

$Cos_{em}(m_i, m_j)$ = Cosine similarity between 10-dimensional emotion reflected by m_i and m_j

$Cos_{prsn}(m_i, m_j)$ = Cosine similarity between 35-dimensional personality reflected by m_i and m_j

(Calculation of topic, emotion, and personality is described in Section 3.2.2)

3.4 Summary

The objective of this research is to design retweet prediction model using users' behavior related features. Third party lexicons are used to mine users' behavioral pattern. This chapter gives a detailed description of features which are used to represent user behavior and the steps we follow to extract them from users' posts. The architectures of proposed models give a thorough description of different stages of work which have been followed to design both machine learning and matrix factorization retweet prediction models. In case of machine learning models, performance of models with only tweet, only retweets, and tweet-plus-retweet based profile are checked. The architecture of models with different profile is same, the only difference is in the process of their profile generation. In case of matrix factorization model, the architecture and algorithm of two models using approach 1 and approach 2 are similar. The only difference between these approaches is the method which is used to regularize the message similarity in the latent and observed space.

Chapter 4

Experiment and Result Analysis

This chapter describes the design of our experiments which is followed by discussion on results achieved from the experiments. Description on the experiment design includes the details of our data collection process and construction of our retweet prediction models.

4.1 Experiment Design

Our objective is to build retweet prediction model based on features related to users' behavior using both machine learning and matrix factorization methods and compare the performance of the proposed model with baseline models. The experiment includes the following steps: data collection, data pre-processing, feature generation, model design, and evaluation. In the data collection step, Twitter API was used to collect data from Twitter. Twitter data was then stored in a database. In the data preprocessing step, the downloaded tweets were preprocessed based on the requirement of the planned feature extraction strategy. Original texts were used for all types of feature extraction. Hashtags were only segmented for personality calculation, and kept as they were for other steps. The feature generation and model design steps were different for machine learning model and matrix factorization model. In case of machine learning model, a user's profile was generated based on their past posts (tweets, retweets, or both) and tweet profile was generated based on the text of the tweet. User profile or target tweet profile was represented by selected content-based explicit and implicit features. Similarity measures were used to calculate similarity between user profile and target tweet profile. These similarity scores were used as features fed to machine learning methods. In case of matrix factorization model, user

profile was not required. Only the retweet matrix (whether a user retweets a tweet) is needed, together with similarity scores between tweets to constrain the objective function. In this phase, tweet profiles were calculated, which include content-based explicit and implicit features. They were then used for similarity calculation. The evaluation of matrix factorization model was done using data generated in reconstructed user-message retweet matrix.

The whole experiment was divided in three stages. In the first stage, we implemented the proposed machine learning retweet prediction models, evaluated the performances of the models and compared their performance with baseline models. The baseline models were developed using explicit and implicit features used in previous research works. The first baseline model (F-UUH) is based on the following explicit features: user-mention, hashtag, and URL. Jaccard similarities between user-mention, hashtag, URL based user profile and target tweet profile were used as input features to the machine learning model. These explicit features had been used successfully by previous researchers (Peng et al., 2011; Suh et al., 2010; Xu & Yang, 2012; Naveed et al., 2011; Uysal & Croft, 2011; Kim & Yoo, 2012). The second baseline model (F-UUH_TF-IDF) used user-mention, hashtag, and URL as explicit features; and Term-Frequency-Inverse-Document-Frequency (TF-IDF) score as the implicit feature. Jaccard coefficient is used to calculate the similarity between user-mention, hashtag, URL based user profile and target tweet profile whereas cosine similarity measure is used to calculate the similarity between TF-IDF based user interest vector (calculated based on past posts) and target tweet vector. TF-IDF technique is investigated in a number of state-of-the-art works (Chen et al., 2010; Yang et al., 2010) and is a well-accepted baseline strategy (Huang et al.,2014; Lu et al.,2012) for retweet prediction. The third baseline model was developed using user-mention, hashtag, URL as explicit features and topics extracted using general Latent Dirichlet allocation (LDA) as the implicit feature. Jaccard coefficient is used to calculate the similarity between user-mention, hashtag, URL based user profile and target tweet profile whereas cosine similarity measure is used to calculate the similarity between LDA based topic vector (calculated based on past posts) and target tweet topic vector. LDA is popular topic modeling technique used to model the topic of tweets (Naveed et al., 2011; Xu & Yang, 2012; Roberts et al., 2012; Hong et al., 2011). In the second stage, we implemented matrix factorization retweet prediction models with proposed regularization terms, evaluated their performances and compared their performance with baseline

matrix factorization model (Koren et al., 2009). In the final stage, we compared the performance of the best-performing machine leaning model and the best-performing matrix factorization model to check which one performs better in developing retweet prediction model.

We used Python programming language to write all the programs used in this research. We used personal computer with the following specification: Intel(R) core(TM) i5-8250U CPU @ 1.60GHz, 8GB RAM, and 64-bit operating system. Some of our experiments were done using facilities provided by the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharc net.ca) and Compute/Calcul Canada. Sharcnet provides high performance computing support for the researchers in Canada through different system, software, and storage support. We used their system "Iqaluk" which is suitable for large jobs.

4.2 Collection of Data

Data collection is an important step for retweet prediction research. Some research used publicly available dataset (Jiang et al., 2015; Naveed et al., 2011; Wang et al., 2015) whereas some collected data from Twitter on their own. For this research, we have collected our own dataset using Twitter Application Programming Interfaces (API). Twitter provides APIs to the developers to get access to Twitter social network data. Researchers use these APIs to gather required information regarding users, their networks, and tweets (Uysal & Croft, 2011; Peng et al., 2011; Lu et al., 2012; Xu et al., 2012; Zhang et al., 2016; Suh et al., 2010; Petrovic et al., 2011; Sun et al., 2013; Luo et al., 2013; Lee et al. 2015; Can et al., 2013). Twitter offers two types of APIs: REST and Streaming. These APIs provide different methods to get data such as user status information, users' tweets, and users' follower/followee information. REST API allows developers to get information based on specific parameters whereas Streaming API delivers live tweet data based on query. These APIs are available through Oauth-based authorization system. Researchers chose API based on the requirement for their research. In this Oauth-based authorization system, researchers create an app in Twitter using their Twitter account. Twitter then provides four types of keys for the created application: consumer key, consumer secret, access key, and access secret. These keys are used to communicate with Twitter

API for developing the application. Both streaming and REST API were used for data collection part of this research. Initially streaming API was used to find active users. Streaming API was used for two days, filtering with hashtags which include '#machinelearning', '#datascience', '#python', '#TorontoMapleLeafs', '#abuse', '#attack', '#peace', '#BlackFriday', '#party', '#BlueJays', '#iPhoneX'. We chose a variety of topics as hashtags, including a few hot topics of that time. The purpose of this filtering process is to cover a variety of users who have interests in different topics and potentially have different behaviors and different retweet decision making processes. Then we checked users' posts in three consecutive days and selected users who have at least five posts. Using REST API, we downloaded three months' data from around 2800 users (from September 06, 2017 to December 06, 2017), with a total of 4179367 posts where number of retweet was 2616063 and rest were tweets. Then we kept data from 1136 users whose numbers of posts in these three months range from 90 to 3000. We consider that a range from 30-1000 per month is a reasonable number for an active user. The distribution is longtail where majority of users have had about 100-300 posts per month. We have also downloaded posts from users' followees to get negative retweet data. Negative retweets are the tweets which appear in a user's timeline but are not retweeted by them. In total, we worked with around 1.6 million posts. Our downloaded dataset contains the following information for every user: a user's screen name, account creation time, number of status counts, number of followers, number of friends, etc. For every post, the dataset has the following information: tweet identification number, posting time, content information such as hashtag, user-mentions, and URLs. If posting is a retweet, then retweet identification number, author identification number, author's friends count, author's follower count, and author's status count are included. We did not use all of the downloaded information in this research. They are saved for future work. We saved the information in MySQL database.

4.3 Implementation of Machine Learning Based Retweet Prediction Model

To build machine learning based retweet prediction model, we used three months' data from 1136 users (from September 06, 2017 to December 06, 2017). For this design, we need users' past data to create their profiles which are then used to predict their future preference.

Every user's first two months' (September 06, 2017 – November 06, 2017) data was used to create user profile and retweets from last month's (November 06, 2017 - December 06, 2017) data was used as positive examples to train the retweet prediction model. Total 1006781 posts were processed to create user profile, and 308593 retweets were used as positive examples, 308593 non-retweets were used as negative examples for evaluating the model.

Since retweet prediction is considered as a binary classification problem, it classifies a target tweet as positive (being retweeted) or negative (not being retweeted). Therefore, along with positive examples, we need negative examples to train the prediction model. Negative examples are the tweets that are tweeted by a user's followees, so appear in user's timeline but not retweeted by the user. To get negative data, for every user, we selected their followees who were retweeted by the user in the last month (November 06, 2017 - December 06, 2017) and took all tweets posted that month. As negative examples we kept non-retweeted posts from these followees. Selection of negative example is a tricky task because there can be several reasons for which a user does not retweet a tweet, for example, he does not notice the tweet, he is no longer interested in the author's posts, or he does not like or support the information expressed by the tweet. Our purpose is to find tweets which he has seen but did not retweet. For every user, the number of negative retweets is much higher than that of positive ones. To create a balanced dataset, we randomly sampled negative retweets, making sure the size of the negative sample set is same as the size of the positive sample set. We also checked the performance of the model with two imbalanced dataset. Firstly, we checked the performance of the model with imbalanced dataset where the number of non-retweets is twice as much as the number of retweets. We chose this ratio because in real life scenario the number of available non-retweets is much larger than the number of available retweets. Secondly, we checked the performance with another imbalanced dataset where the number of retweets was twice as much as the number of non-retweets. We chose this imbalance ratio because for information retrieval problems positive examples give more consistent and effective training feedback (Manning et al., 2008).

We also wanted to explore whether user's past retweets or past tweets provide more information for retweet prediction. We built three retweet prediction models. The first model follows conventional strategy which used user's past tweets and retweets to create user profile;

the second one used user's past retweets only to create user profile and the third one used user's past tweets only to create user profile. The second model was developed to investigate whether user's past retweets have better potential in prediction of future retweets than past tweets or both tweets-plus-retweets. The third model explores the performance of user's own tweets to predict his future retweet preference.

Since retweet prediction is considered as a classification problem which aims to predict a target tweet as positive or negative, the proposed retweet prediction models were built using python scikit-learn machine learning package which provides machine learning tools for classification, regression, and clustering problems. Scikit-learn has implementation of different classification algorithms such as Decision Tree, Support Vector Machine, Linear Discriminant Analysis, Naïve Bayes, Random Forest, etc. For this research, we used XGBoost and Random Forest to build retweet prediction models. The parameters of machine learning methods are fine tuned to create the model. Our purpose of using two methods is to check whether there is any significant difference in performance when using two different machine learning techniques.

The performance of a retweet prediction model depends on its effectiveness on predicting the class labels of unlabeled tweets. Here the class labels of tweets are either positive or negative. To measure the performance, we arranged the data (positive and negative retweets) in temporal order and selected the initial 80% of the data for training and the remaining 20% of the data for testing purpose. We also used time-split 10-fold cross validation technique[14] on the dataset to

Table 4.1: Definition of different models

Mode Index	**Feature set definition**
F-full	$f_{um}, f_{url}, f_{ht}, f_{tp}, f_{em}, f_{per}$
F-TEP	f_{tp}, f_{em}, f_{per}
F-TP	f_{tp}, f_{per}
F-TE	f_{tp}, f_{em}
F-EP	f_{em}, f_{per}

[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit

measure the performance of our models. We also checked the performances of the models using randomly selected training and testing data and got similar results. Results from random selection (80/20 split and 10-fold cross validation) are given in appendix A (Table A.2 and A.3). The intention is to train the model with samples where the class labels are known (given) and then use the trained model to predict the class labels of samples (test data) where the class labels are unknown (not given). Since our objective was to explore the combination effects of different implicit features on future retweet decision, we implemented retweet prediction models with different combinations of the feature sets. The definitions of different models are given in Table 4.1. As shown in Table 4.1, model index is relevant to the features included in the model. F-full is the model which includes all 6 features; F-TEP is the model which includes topic, emotion, and personality related features; F-TP is the model which includes topic and personality related features; F-TE is the model which includes topic and emotion related features; and F-EP is the model which includes emotion and personality related features. Features are named as $f_{um}, f_{url}, f_{ht}, f_{tp}, f_{em},$ and f_{per} respectively. Descriptions of these features are given in Table 3.1.

4.4 Implementation of Matrix Factorization Based Retweet Prediction Model

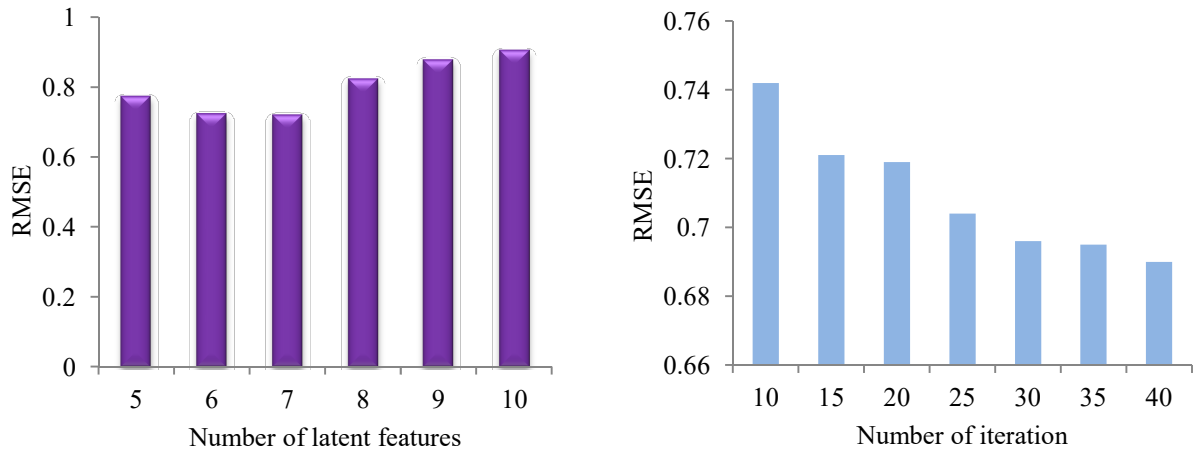
For matrix factorization retweet prediction model, we need to design a user-message retweet matrix where value of each entry represents whether the user retweets the message (positive retweet), whether the user see the message in the timeline but do not retweet it (negative retweet), or whether there is no relation between the user and the message. For this model, if we take all positive and negative retweet for every user, the matrix would become too large and sparse. Experiment with this kind of huge matrix would often cause out-of-memory error. So, to reduce the sparsity as well as to make the size of the matrix manageable during the experiment, we selected a sample of messages from all positive and negative retweet messages. The sample of messages is selected in such a manner that each column of the matrix has more than one non-zero value. So, our user-message retweet matrix is turned into an 1124 × 60348 matrix where 1124 is the number of users and 60348 is the number of messages (positive and negative retweets).

We developed matrix factorization retweet prediction models using programming language Python. We arranged the data in temporal order and used the initial 80% of the data to train the model, and the rest 20% of the data for testing. We have also checked the performance of he models with randomly selected training and testing data. For experiment, entries in user-message retweet matrix R are set in the following manner: 2 for positive retweet, i.e. user retweets the message; 1 for negative retweet, i.e. user did not retweet the message after seeing it; 0 if there is no relation between user and message, i.e. user does not follow the author of the message; test entries are set to 0.

In matrix factorization technique, an important step was to select the number of latent features and number of iterations. Higher number of latent features usually gave better accuracy. But we found that number of iterations had to increase largely to get the benefit of increased number of latent features. For example, with 7 latent features and 20 iterations the Root Mean Square Error (RMSE) is around 0.71, but with 8 latent features, it takes 45 iterations to achieve RMSE close to 0.71. A large number of iterations were very time and resource consuming. So, for number of latent features and iterations, we kept a number which was reasonable in terms of performance and time complexity. Performance of proposed retweet prediction model (with approach 1) with different number of latent features and different number of iterations are shown in Figure 4.1. At this stage, as a measure of performance, we calculated Root Mean Square Error (RMSE) of the prediction model. Root Mean Square Error (RMSE) is a measure which finds the difference between actual value and predicted value. In this case, errors are squared and averaged before taking their square root. It gives more weight to the large errors (Chai & Draxler, 2014). Smaller RMSE indicates better accuracy. RMSE can be calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(A_n - P_n)^2}{N}} \tag{4.1}$$

where N is the number of instances in the dataset; A_n is the actual and P_n is the predicted value for n^{th} instance.



(a) (b)

Figure 4.1: Performance of matrix factorization (with approach 1) based retweet prediction model with (a) different number of latent feature and (b) different number of iterations

We presented the result with the number of latent features starting from 5. We also tried with number of latent features less than 5 and did not get satisfactory result. The reason can be that number of latent features less than 5 is too low to take the full advantage of latent feature generation process of matrix factorization technique. As presented in Figure 4.1, when the number of latent features increases from 5 to 6 and then from 6 to 7, RMSE values reduces from 0.8 to 0.7. But after 7, increasing the number of latent features deteriorates the performance (RMSE increases). So, we decided to use 7 latent features. We also need to identify a reasonable value for number of iterations. We checked the performance of the model with increasing number of iterations starting from 10. We found that up to 30 iterations, there is a noticeable improvement in terms of RMSE values. After 30, the improvement is very small and it also takes very long time to finish the computation process. So, we decided to use 30 iterations to keep a reasonable computational time. Performance of retweet prediction model with approach 2 was checked using the same number of latent features and iterations, making sure the results of the two models are comparable.

The trade-off parameters β, γ, α_1, and α_2 regulate the effects of different regularization terms on objective function of matrix factorization. We checked different combinations of β, γ, α_1 to

find the one for which the matrix factorization model (with approach 1) gave the best performance (lowest RMSE). Some of these combinations are listed in Table 4.2. The selected combinations (highlighted in Table 4.2) might not be the perfect one, but they give satisfactory performance for our experiment. For matrix factorization model with approach 2, we kept the same β, γ as selected for approach 1 to keep these approaches comparable. Since α_1 and α_2 are two different trade-off parameters used by different regularization terms in approach 1 and approach 2 respectively, so for approach 2, we tuned the value for α_2. We checked the performance of matrix factorization model (with approach 2) with different α_2 values and found that for α_2 =0.8, it gave the best performance. We also used another parameter δ for calculating similarity between messages which was used in our proposed new regularization terms (see equation 3.24). Here, δ is a constant which controls the contribution of explicit and implicit features in the calculation of final similarity between messages. We checked with different values ranging from 0.1 to 0.9 and found that with $\delta = 0.2$, we are getting the best performance for proposed models.

Table 4.2: Performance with different trade-off parameters

β	γ	**α_1**	**RMSE**
0.01	0.0001	0.00001	0.734
0.001	0.00001	0.9	0.819
0.01	**0.001**	**0.000001**	**0.713**
0.01	0.001	0.005	0.737

4.5 Performance Evaluation

Performance evaluation of a model is an important step of designing a prediction model which reports the effectiveness of the model in predicting class labels of future unseen data. To

evaluate the performance of both machine learning and matrix factorization retweet prediction models, we used three metrics: precision, recall, and F1-score. Since retweet prediction is a binary classification problem, we describe the metrics for binary classification where the positive retweets represent relevant instances and the negative retweets represent the opposite. Precision reports the fraction of retweets classified as positive are truly positive retweets. Recall refers to the fraction of the positive retweets that are correctly classified as positive retweets. F1-score gives the average of precision and recall value. Precision, recall, and F1-score are defined in equation 4.2 4.3, and 4.4 respectively. Each of these metrics has its unique significance. In retweet prediction problem, prediction of positive retweets is more important than prediction of negative ones because finding the correct target messages for target user holds more significance than filtering out the negative instances. We used precision to show a model's performance in predicting positive retweets correctly out of all its positive predictions and recall to define a model's performance in predicting positive retweets correctly out of all the positive retweets. Precision may not give a good idea about a model's performance in predicting all retweets correctly; on the other hand, recall may not give a good idea about a model's behavior in predicting non-retweets as retweets. Therefore, we used both metrics to show a model's performance. We also used F1-score to define performance of the model because with the harmonic mean of precision and recall, it shows the balance between them. We wanted to propose a balanced model which gives a reasonable performance in terms of all the three metrics.

$$Precision = \frac{t_p}{t_p + f_p} \tag{4.2}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{4.3}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{4.4}$$

where, t_p = number of positive retweets classified as positive

t_n = number of negative retweets classified as negative

f_p = number of negative retweets classified as positive

f_n = number of positive retweets classified as negative

For machine learning based model, scikit-learn package provides implementation of our required evaluation metrics (F1-score, precision, and recall). In case of matrix factorization models, we did our own calculation using equation 4.2, 4.3, and 4.4. For this purpose, from reconstructed retweet matrix we calculated the median of the generated value. Then values that are greater than the median are considered as positive prediction and the rest are considered as negative prediction.

4.6 Result Analysis of Machine Learning Based Models

For machine learning based models, we developed several different feature sets (Table 4.1) for retweet prediction. These feature sets were used to generate XGBoost and random forest models which in turn were tested to find the performance of the designed model. Both XGBoost and random forest models gave similar result while XGBoost models showed a little better performance than random forest models. Our objectives were to explore the performance of our proposed models and compare their performances with baseline models. Five different retweet prediction models were developed using the proposed five feature sets (as defined in Table 4.1). Users' profiles were created based on their past tweets-plus-retweets, Precision, Recall, and F1-score of these models using XGBoost and random forest methods are presented in Table 4.3 and 4.4. In these tables, we presented the results for experiments done using balanced dataset (numbers of positive and negative examples are the same); and 80% data for training and 20% data for testing purpose. We also did time-split 10-fold cross validation on the dataset and got similar results. Here, we report the results from 80/20 split. Results from time-split 10-fold cross validation is given in Appendix A Table A.1

Table 4.3: Performance of baseline and proposed machine learning based models using XGBoost method

	Models	Precision	Recall	F1-score
Baseline Models	F-UUH	**0.7520**	0.5239	0.6172
	F-UUH_TFIDF	0.7480	0.5559	0.6378
	F-UUH_LDA	0.7508	0.5590	0.6409
Proposed Models	F-Full	0.7380	0.6136	**0.6701**
	F-TEP	0.5489	0.6472	0.5940
	F-TE	0.5453	0.5963	0.5697
	F-TP	0.5221	**0.6547**	0.5810
	F-EP	0.5499	0.6312	0.5877

Table 4.4: Performance of baseline and proposed machine learning based models using random forest method

	Models	Precision	Recall	F1-score
Baseline Models	F-UUH	0.7525	0.5282	0.6207
	F-UUH_TFIDF	0.7431	0.5377	0.6239
	F-UUH_LDA	0.7487	0.5473	0.6323
Proposed Models	F-full	**0.7532**	0.5901	**0.6636**
	F-TEP	0.5478	**0.6369**	0.5890
	F-TE	0.5434	0.5710	0.5568
	F-TP	0.5214	0.6361	0.5731
	F-EP	0.5495	0.6263	0.5854

In Table 4.3 and 4.4, we present the performance of our proposed models as well as baseline models using XGBoost and random forest methods respectively. Our proposed model F-full considers both explicit and implicit features whereas the other proposed models (F-TEM, F-TE, F-TP, F-EP) are based on the combination of different implicit features. Among the proposed models, performance of F-full is the best in terms of precision. However, it is slightly worse than the baseline models when XGboost is used. When random forest is used, its performance is better than the baselines. In terms of recall, performance of all proposed models is better than the baseline models. Recall of F-TEP and F-TP are better than other proposed models. Recall of F-full with XGBoost is better than recall of F-full with random forest. By checking the performance of the proposed models developed using different combinations of implicit content features (F-TEP, F-TP, F-TE, F-EP), it is evident that implicit features only are not good for problems where the requirement is to have higher precision. Therefore, it can be said that it needs both explicit and implicit content features to design a balanced model (model good for both precision and recall). Using both XGBoost and random forest, as per F1-score, performance of F-full is the best among all proposed models, and it is also better than the F1-score of baseline models. Therefore, we are considering proposed model F-full as the best overall model.

Using both XGBoost and random forest methods, the improvement from F-full is 3%-6% when compared with F1-scores of baseline models F-UUH, F-UUH_TFIDF, and F-UUH_LDA. Tables 4.3 and 4.4 show that in terms of precision, baseline models are better than proposed models (F-TEP, F-TP, F-TE, F-EP). In terms of recall, all proposed models are better than baseline models. The proposed model F-full can be considered as the most preferred model because it gives precision comparable to other baseline models while provides recall better than other baseline models. For many application of retweet prediction, recall is more important than precision because recall measures a model's performance in predicting as many as possible positive retweets whereas precision is concerned about a model's prediction quality. The result shows that models using combination of implicit features could be a good option when the objective is mostly to find out potential retweets or retweeters as many as possible.

Baseline model F-UUH_LDA used general LDA to extract topics from tweets whereas proposed models used Twitter-LDA to extract topics from tweets. To compare the performance of general LDA and Twitter-LDA, baseline model F-UUH_LDA was reconstructed using Twitter LDA. The F1-score of the reconstructed F-UUH_LDA (using Twitter LDA) is 2% higher than the F1-score of baseline F-UUH_LDA (using general LDA) model. Better performance of reconstructed F-UUH_LDA justifies our selection of Twitter-LDA over general LDA for designing the proposed models.

Table 4.5: Performance of F-full using different profile

Profile	Precision	Recall	F1-score
Tweet-plus-Retweet	0.7380	0.6136	0.6701
Retweet-only	0.6938	0.6429	0.6674
Tweet-only	0.5873	0.5944	0.5757

Another objective of this research was to explore whether a user's tweet-plus-retweet, retweet-only or tweet-only profile provide more information for their future retweet prediction. Table 4.5 presents the performance of best performed proposed model F-full developed using three types of profiles. We examined the performance of F-full with different profiles using both XGBoost and random forest methods, and got similar result. In table 4.5, we present the results achieved using XGBoost method. The results show that in terms of F1-score, performance of F-full with tweet-plus-retweet profile is slightly better than its performance with retweet-only profile. Performance of F-full with tweet-only profile is worst among the three types of profiles. According to the results from the t-test, the difference between the performance (in terms of F1-score) of tweet-plus-retweet profile and retweet-only profile is not significant because the *p*-value is 0.292893 (the difference is not considered significant when $p > 0.05$). On the other hand, the difference between the performance of tweet-plus-retweet profile and tweet-only profile is significant because the *p*-value is 0.036454 (the difference is considered significant when $p <$

0.05). The findings suggest that retweet-only profile can be an alternative of tweet-plus-retweet profile, and users' past tweets do not provide much information for their future retweet prediction whereas past retweets are main contributor of information. In terms of recall, retweet-only profile gives better performance than tweet-plus-retweet profile. So, if the requirement of the problem is to have higher recall or lower processing time, then retweet-only profile is a better choice because it can give better recall as well as require less processing time to create user profile. Figure 4.2, Figure 4.3, and Figure 4.4 show the visualizations of retweet only, tweet only, and tweet-plus-retweet based classification report with best-performing feature set F-full for positive (represented by 1) and negative classes (represented by 0). This report can be used to select a model that has redder metrics, i.e., have stronger and more balanced classification capability. These visualizations show the classification reports on per-class basis. Classification reports with retweet only and twee-plus-retweet profile have similar redder metrics for predicting positive retweets whereas tweet only profile shows worse result. As per the report, tweet-plus-retweet profile outperforms retweet only profile in terms of recall for predicting negative retweets. Since prediction of negative retweet is less important for retweet prediction problem, we can consider retweet only profile as a comparable alternative to tweet-plus-retweet profile.

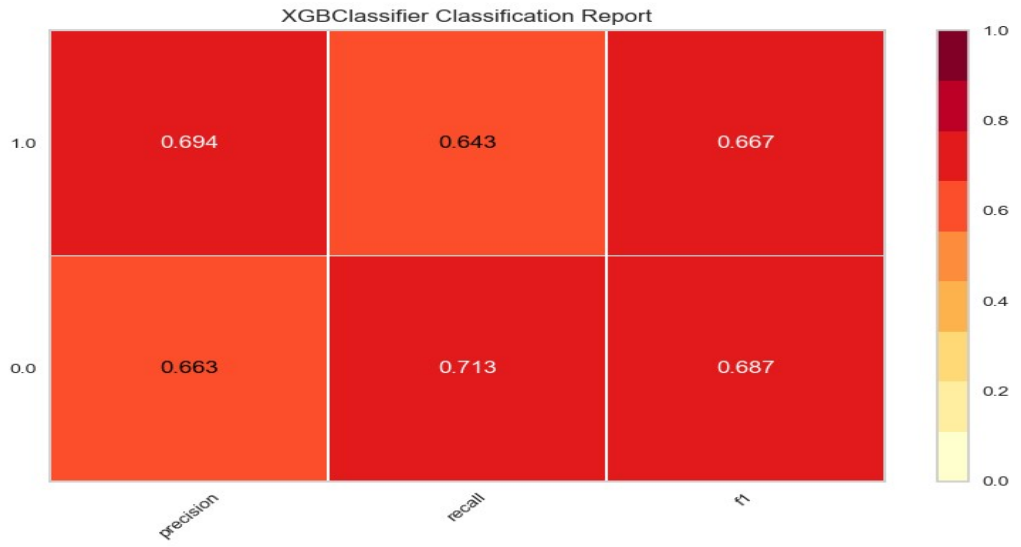


Figure 4.2: Classification report of model F-full based on retweet only profile using XGBoost method

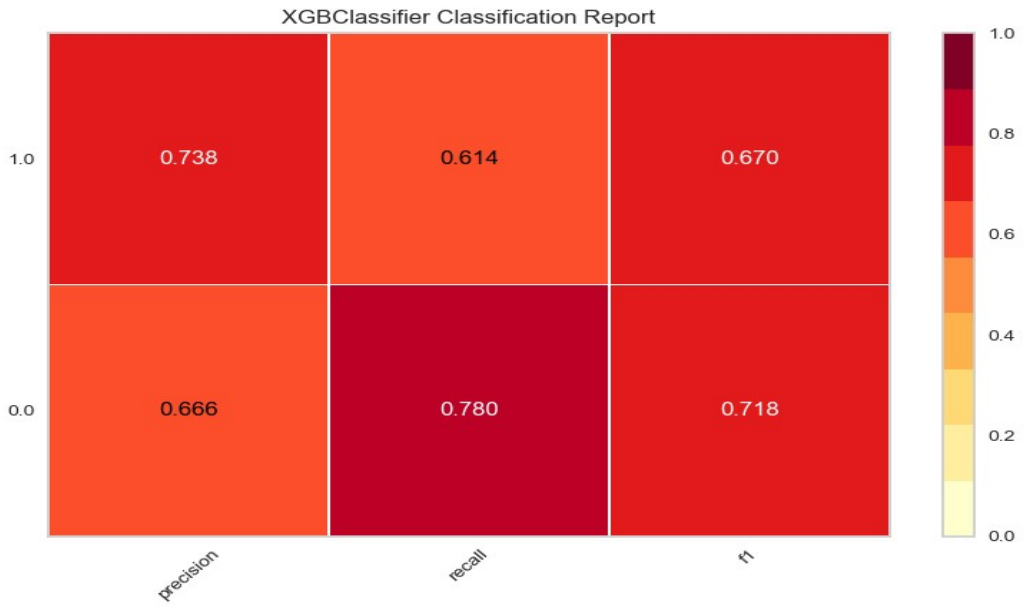


Figure 4.3: Classification report of model F-full based on tweet-plus-retweet profile using XGBoost method

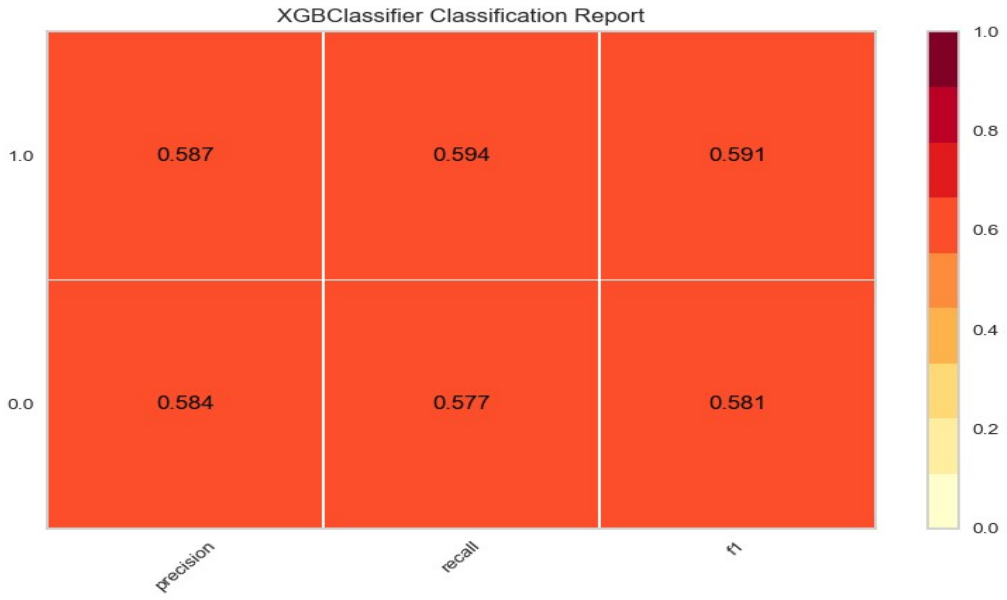


Figure 4.4: Classification report of model F-full based on tweet only profile using XGBoost method

Table 4.6: Performance of F-full using balanced and imbalanced dataset

Ratio [retweet: non-retweet]	Precision	Recall	F1-score
1:1	0.7380	0.6136	0.6701
1:2	0.6114	0.5654	0.5875
2:1	0.7322	0.8582	0.7900

We also tested the performance of the best-performing proposed model (F-full) using imbalanced dataset. Table 4.6 presents the results of F-full with balanced and imbalanced datasets using XGBoost method. Along with balanced dataset (equal number of retweets and non-retweets), we examined the model with two imbalanced datasets. In one imbalanced dataset, the number of non-retweets is twice the number of retweets; and in the other imbalanced dataset, the number of retweets is twice the number of non-retweets. As per the results of imbalanced dataset, performance of the model with 2:1 ratio is better than performance of the model with ratio 1:2 for predicting retweet (positive instance). The results are reasonable because in case of ratio 2:1, model is trained with more retweets than non-retweets, hence its performance for predicting retweet is better. In case of ratio 1:2, the model is trained with more non-retweets than retweets, so its performance for predicting retweet is lower than others. The ratio of 1:2 is closer to the real scenario because usually there are more non-retweets than tweets.

From the first step of our experiment, we got a conclusion that model F-full which is based on both explicit and newly proposed implicit features performs better than baseline models. It is also evident that users' past retweets could provide more information for predicting future retweets compared to their past tweets. Performance of retweet only profile is comparable to performance of model with tweet-plus-retweet profile. Moreover, it gives the advantage of lower processing time than tweet-plus-retweet profile for creating user profile.

4.7 Result Analysis of Matrix Factorization Based Models

We proposed matrix factorization retweet prediction models with new regularization terms based on similarity between a user's messages. Two different approaches are proposed for similarity-based regularizers. The performances of the proposed approaches are compared with the performance of the baseline matrix factorization approach. Table 4.7 shows the performance of matrix factorization retweet prediction models with approach 1, approach 2 and baseline when data was arranged in temporal order and initial 80% data was used for training and the rest 20% data was used for testing. We have also checked the performance of the model with randomly selected training and testing data. Results from random selection is given in Appendix A (Table A.4)

Table 4.7: Precision, recall, and F1-score of matrix factorization retweet prediction models developed using proposed approach 1, approach 2, and baseline.

	Precision	**Recall**	**F1-score**
Baseline	0.6108	0.6426	0.6263
Approach 1	0.6041	0.7094	0.6526
Approach 2	0.6106	0.7233	0.6622

Table 4.7 shows that performance of approach 1 and approach 2 are better than baseline in terms of recall, and F1-score, which explains the benefit of using new proposed regularization terms. In case of precision, performance of approach 1 and approach 2 are comparable to baseline. Recall for approach 2 gave 2% improvement than approach 1, and they both improve the recall value from the baseline model. Performance of approach 2 is the best when compared with approach 1 and baseline in terms of evaluation metrics. Approach 1 improves 3% from baseline in terms of F1-score, 6% in terms of recall. Approach 2 improves 4% from the baseline in terms of F1-score, 8% in terms of recall. It shows that proposed approach 2 is more effective than the proposed approach 1 for designing retweet prediction model.

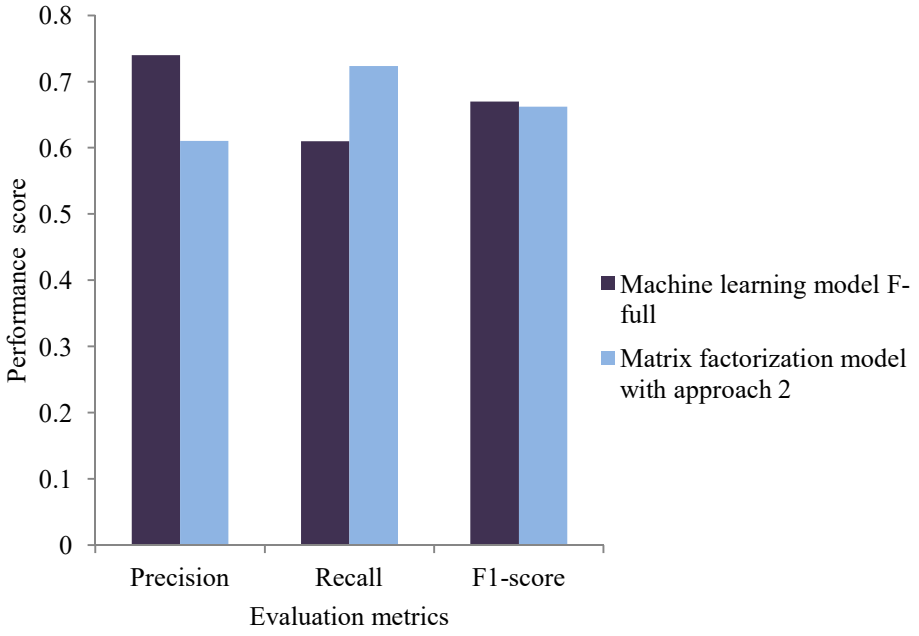


Figure 4.5: Performance of machine learning and matrix factorization retweet prediction models

4.8 Performance Comparison Between Machine Learning Based Model and Matrix Factorization Based Model

Our objective in this experiment was to compare the performance of retweet prediction models using machine learning approach and matrix factorization approach. We wanted to compare the performance of human extracted features representing users' behavioural patterns as done in machine learning model with matrix factorization model which makes retweet decision based on machine extracted latent features. We compare the best-performing machine learning model (Model F-full using XGBoost method) with best-performing matrix factorization model (approach 2) in Figure 4.5. From the figure, we can see that none of the models outperformed the other in terms of all three evaluation metrics (precision, recall, and F1-score). Machine learning model performed better than matrix factorization model in terms of precision; matrix factorization model performs better than machine learning model in terms of recall. In terms of

precision and F1-score, machine learning model is 12% and 1% better respectively than matrix factorization model. In terms of recall, approach 2 of matrix factorization model is about 11% better than machine learning model. Since there is no clear winner, the preferred technique should be selected based on the requirement of the problem. If a model's performance quality measured by precision (i.e. prediction of positive retweets as positive) is more important, machine learning model is a better choice. Matrix factorization model can be a better choice when a model's capability of identifying as many positive retweets as possible (as given by recall) is more important. In retweet prediction research, recall is important for marketing campaign (to find as many retweets as possible), identifying all potential retweeters, finding all potential paths for information propagation, etc.; and precision is important for tweet recommendation based on probability of being retweeted, or tweet re-ranking, etc.

Chapter 5

Conclusion and Future Work

Retweet prediction is an important area of research. Many works have been done in this area. The complexity and dynamism of the user behavior and intention within the fast-growing social network site has made it an interesting area of research. Since users are the main actors in retweeting the posts, in this research we wanted to explore their behavioural pattern in case of making retweet decision. We believe that a user's behavior is the main and subtle influencing factor which drives him to take any action. The findings of this research are as follows:

- We explored the impact of users' behavior on their retweet decision. A user's behavior was represented by their interests and attitudes. We used different explicit and implicit content features to represent their interests and attitudes. We found that our proposed model using both explicit (user-mention, hashtag, URL) and implicit content features (topic, emotion, and personality) performed better than baseline models.

- We used a more complete list of emotion feature as implicit content features to develop retweet prediction model. Most of the previous researchers used only positive and negative sentiment to the best of our knowledge, whereas we included 10-dimentional emotion and sentiment which includes anger, anticipation, disgust, fear, joy, sadness, surprise, and trust along with positive and negative sentiment. Past research explored the influence of sentiment to find the retweetability of tweets from global perspective whereas we explored the influence of emotion and sentiment from the perspective of individual user for predicting a user's retweet decision. Our research showed that

emotion and sentiment along with a user's interest-based feature (topic) are good informative features to predict their future decision.

- We introduced 35-dimensional personality as an influential implicit feature to predict a user's retweet decision. Past research used personality along with other content-based features to find retweeters. We showed that personality reflected by the tweets along with other content-based feature has good capability to predict users' retweets.

- We showed the impact of different combinations of implicit content features on users' retweet decision. We found that the combination of topic, emotion-sentiment, and personality and combination of topic and personality gave the best prediction result when using only the implicit content features.

- We found that users behave differently as an author and as a retweeter. We tested the performance of the prediction model when using different user profiling strategies –user profile based on both tweets and retweets, past retweets only, or past tweets only, to find out which strategy provides more information for user's future retweet decision. We found that performance of users' past retweet based profile is comparable to past tweet-plus-retweet based profile. Retweet based profile give better recall than tweet-plus-retweet profile. Compared to conventional tweet-plus-retweet profile, retweet only profile can be used when higher recall is the requirement of the application. Retweet based profile also give the advantage of less processing time, because processing past retweets takes less time than processing both past tweets and past retweets to create user profile. We also found that, past tweets do not have better retweet prediction capability when compared with performance of past retweets. So, different prediction quality from different profiles (tweet-only, retweet-only, and tweet-plus-retweet) shows that a user's behavior is different as an author and as a retweeter.

- We explored the performance of matrix factorization retweet prediction model. We showed that the proposed model that uses user-specific message-similarity-based regularizer performs better than the baseline matrix factorization model for predicting retweet decision.

- We compared the performance of machine learning model with matrix factorization model. We found that none of the models over-performs the other in terms of all performance metrics. So, selection of method would be based on requirement of the task at hand. Machine learning model is preferable when a model's performance quality in important. Matrix factorization can be a preferred choice when model's positive retweet prediction capability is more important.

The major contributions of this research are highlighted below:

- We found that proposed model with implicit content features which represent users' behavioral pattern show good improvement (3%-6%) over baseline models for developing machine learning based retweet prediction model.

- We showed that users behave differently as an author and retweeter. User's retweet-only profile gave performance comparable to performance of conventional tweet-plus-retweet profile. Retweet-only profile performed fairly better (9% in terms of F1-score) than tweet-only profile for retweet prediction problem.

- We showed that matrix factorization retweet prediction model that uses user-specific message similarity for regularizing the objective function, performs better than the baseline matrix factorization model for predicting retweet decision.

The limitations of this research work are as follows:

- We observed that machine learning based model required less computational time than matrix factorization model. But a detailed comparison on their running time is not made in this work. Also, we didn't use the distributed computing platform such as Hadoop to improve the running speed.

- To compute emotion/sentiment and personality features, we relied on previously published database as they are well established and successfully used in past research.

- For baseline models, we could not implement models published in recent years because of the lack of implementation details in the papers.

In future work, we would like to include the following tasks:

- In this research, we explored only topic, emotion/sentiment, and personality to represent a user's behavior. There are still many unexplored implicit content features such as values, beliefs, views on topics which might have impact on user's retweet decision. In future work, we would like to investigate these unexplored latent factors. Also, some users' interest is static whereas for other users, it changes frequently. We would like to identify the temporal pattern of user's changing behavior and study how this kind of behavior drift may affect their retweet decisions. Afterwards, we could incorporate user's dynamic behavior into the prediction model.

- In proposed model, when creating user profile, we did not consider the time decay factor on users' interest. In future, we would like to take this into consideration, e.g., older data has lower weights.

- In our proposed matrix factorization model, recall is better than precision, which signifies that our model's positive retweet prediction capability is better than its overall performance (positive as well as negative retweet prediction capability). In future, we would like to work out a solution to make it more balanced.

- User's fraudulent behavior is a challenging factor when considering retweets as mechanism for information diffusion; especially during emergency situation it can create panic. Retweet fraud can also create false product advertisement which might lead to wrong product review. Researchers are working on this issue (Jiang et al., 2016, Giatsoglou et al., 2015), but retweet prediction research has not yet dealt with user's fraudulent behavior. We would like to tackle this challenge in future.

- We would like to build tweet recommender system for inactive users. Most of the past research did not include inactive users' data because there is not enough data available for inactive users to find their preferences on tweeting/retweeting decision. Researchers also assumed that inclusion of inactive user's data might lower model's accuracy.

However, inactive user can be potential retweeter. A user not posting anything does not really imply that he is not checking Twitter. It might be possible that he is not interested in posting anything or he is not finding anything interesting to tweet/retweet. An effective tweet recommendation according to his interest might be able to make him tweet or retweet more frequently. Inactive user's activeness can be checked from his changing friend list. If he is not posting anything but adding new friends, then it indicates that he is following friends in Twitter. Data from his friends can be good source of information about an inactive user's preference which can be used to recommend tweets for him. Good recommendation may be able to transform a user from inactive to active. Data from third party such as other social networking sites or from his online activity might also be helpful to create inactive user's interest profile.

Appendix A

Additional Experimental Results

Table A.1: Performance of baseline and proposed machine learning based models with time-split 10-fold cross validation using XGBoost method

	Models	Precision	Recall	F1-score
Baseline Models	F-UUH	0.7390	0.4980	0.5944
	F-UUH_TFIDF	0.7234	0.5306	0.6118
	F-UUH_LDA	0.7323	0.5265	0.6123
Proposed Models	F-Full	0.7240	0.6054	0.6588
	F-TEP	0.5364	0.6795	0.5994
	F-TE	0.5390	0.6354	0.5831
	F-TP	0.5180	0.6615	0.5808
	F-EP	0.5394	0.6332	0.5824

Table A.2: Performance of baseline and proposed machine learning based models with random split for 80% training and 20% testing data using XGBoost method

	Models	Precision	Recall	F1-score
Baseline Models	F-UUH	0.7580	0.5359	0.6379
	F-UUH_TFIDF	0.7599	0.5494	0.6377
	F-UUH_LDA	0.7577	0.5462	0.6348
Proposed Models	F-Full	0.7211	0.6387	0.6774
	F-TEP	0.5546	0.6380	0.5933
	F-TE	0.5486	0.6006	0.5734
	F-TP	0.5294	0.6412	0.5799
	F-EP	0.5497	0.5903	0.5693

Table A.3: Performance of baseline and proposed machine learning based models with random split 10-fold cross validation using XGBoost method

	Models	Precision	Recall	F1-score
Baseline Models	F-UUH	0.7432	0.4839	0.5849
	F-UUH_TFIDF	0.7491	0.5025	0.6007
	F-UUH_LDA	0.7413	0.5237	0.6128
Proposed Models	F-Full	0.7239	0.5970	0.6539
	F-TEP	0.5455	0.6490	0.5926
	F-TE	0.5408	0.6183	0.5767
	F-TP	0.5220	0.6811	0.5907
	F-EP	0.5428	0.6352	0.5853

Table A.4: Precision, recall, and F1-score of matrix factorization retweet prediction models developed using proposed approach 1, approach 2, and baseline using randomly selected 80% data for training and 20% data for testing.

	Precision	Recall	F1-score
Baseline	0.6279	0.8257	0.7136
Approach 1	0.6340	0.8806	0.7372
Approach 2	0.6452	0.8872	0.7471

References

1. Abdullah NA, Nishioka D, Tanaka Y, Murayama Y (2015) User's action and decision making of retweet messages towards Reducing misinformation spread during disaster. Journal of information processing. 2015;23(1):31-40.
2. Acar A, Muraki Y (2011) Twitter for crisis communication: lessons learned from Japan's tsunami disaster. International Journal of Web Based Communities. 2011 Jan 1;7(3):392-402.
3. Adali S, Escriva R, Goldberg MK, Hayvanovych M, Magdon-Ismail M, Szymanski BK, Wallace WA, Williams G (2010) Measuring behavioral trust in social networks. In 2010 IEEE International Conference on Intelligence and Security Informatics 2010 May23 (pp. 150-152). IEEE.
4. Adali S, Golbeck J. Predicting personality with social behavior. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on 2012 Aug 26 (pp. 302-309). IEEE.
5. Adamic LA (1999) The small world web. In Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries 1999 Sep 22 (pp. 443-452). Springer-Verlag.
6. Arora S, Venkataraman V, Donohue S, Biglan KM, Dorsey ER, Little MA. High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on 2014 May 4 (pp. 3641-3644). IEEE.
7. Asur S, Huberman BA (2010) Predicting the future with social media. In: 2010 IEEE/WIC/ACM international conference web intelligence and intelligent agent technology (WI-IAT), vol 1, pp 492–499
8. Berger J (2011) Arousal increases social transmission of information. Psychological science. 2011 Jul;22(7):891-3.
9. Blei DM (2012) Probabilistic topic models. Communications of the ACM. 2012 Apr 1;55(4):77-84.
10. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research. 2003; 3(Jan):993-1022.
11. Boecking B, Hall M, Schneider J (2015) Event prediction with learning algorithmsA study of events surrounding the Egyptian revolution of 2011 on the basis of micro blog data. Policy Internet 7(2):159–84
12. Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1, the center for research in psychophysiology, University of Florida; 1999.

13. Bravo-Marquez F, Frank E, Mohammad SM, Pfahringer B (2016) Determining word-emotion associations from tweets by multi-label classification. In: Proceedings of the 2016 IEEE/WIC/ACM international conference on web intelligence, USA, 2016, IEEE Computer Society, p 536539, https ://doi.org/10.1109/WI.2016.90

14. Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

15. Can EF, Oktay H, Manmatha R (2013) Predicting retweet count using visual cues. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management 2013 Oct 27 (pp. 1481-1484). ACM.

16. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: The million follower fallacy. In ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social 2010.

17. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature (2014) Geoscientific Model Development. 2014 Jun 30; 7(3):1247-50.

18. Chen K, Chen T, Zheng G, Jin O, Yao E, Yu Y (2012) Collaborative personalized tweet recommendation. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval 2012 Aug 12 (pp. 661-670). ACM.

19. Chen J, Nairn R, Nelson L, Bernstein M, Chi E. Short and tweet: experiments on recommending content from information streams. InProceedings of the SIGCHI Conference on Human Factors in Computing Systems 2010 Apr 10 (pp. 1185-1194). ACM.

20. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794). ACM.

21. Comarela G, Crovella M, Almeida V, Benevenuto F (2012). Understanding factors that affect response rates in twitter. In Proceedings of the 23rd ACM conference on Hypertext and social media (pp. 123-132). ACM.

22. Costa Jr PT, McCrae RR. Revised NEO personality inventory (NEO-PI-R) and NEO five-factor (NEO-FFI) inventory professional manual. Odessa, FL: PAR. 1992.

23. Firdaus SN, Ding C, Sadeghian A (2017) Topic Specific Emotion Detection for Retweet Prediction, submitted in International Journal of Machine Learning and Cybernetics, Special Edition : Affective and Sentimental Computing (final revised version submitted)

24. Fire M, Tenenboim-Chekina L, Puzis R, Lesser O, Rokach L, Elovici Y. Computationally efficient link prediction in a variety of social networks. ACM Transactions on Intelligent Systems and Technology (TIST). 2013 Dec 1;5(1):10.

25. Gao S, Ma J, Chen Z (2015) Modeling and predicting retweeting dynamics on microblogging platforms. In Proceedings of the Eighth ACM International Conference on

Web Search and Data Mining 2015 Feb 2 (pp. 107-116). ACM.

26. Golbeck J, Robles C, Edmondson M, Turner K. Predicting personality from twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on 2011 Oct 9 (pp. 149-156). IEEE.
27. Golbeck J, Robles C, Turner K. Predicting personality with social media. In CHI'11 extended abstracts on human factors in computing systems 2011 May 7 (pp. 253-262). ACM.
28. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Alzheimer's Disease Neuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage. 2013 Jan 15;65:167-75.
29. Gupta M, Gao J, Zhai C, Han J (2012) Predicting future popularity trend of events in microblogging platforms. Proc Assoc Inf Sci Technol 49(1):1
30. Hoang TA, Lim EP (2013) Retweeting: An act of viral users, susceptible users, or viral topics? In Proceedings of the 2013 SIAM International Conference on Data Mining 2013 May 2 (pp. 569-577). Society for Industrial and Applied Mathematics.
31. Hong L, Dan O, Davison BD (2011) Predicting popular messages in twitter. In Proceedings of the 20th international conference companion on World Wide Web 2011 Mar 28 (pp. 57-58). ACM.
32. Huang D, Zhou J, Mu D, Yang F (2014) Retweet behavior prediction in twitter. In Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium; 2:30-33. IEEE.
33. Jenders M, Kasneci G, Naumann F (2013) Analyzing and predicting viral tweets. In Proceedings of the 22nd international conference on World Wide Web companion 2013 May 13 (pp. 657-664). International World Wide Web Conferences Steering Committee.
34. Jiang B, Liang J, Sha Y, Wang L (2015) Message clustering based matrix factorization model for retweeting behavior prediction. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management 2015 Oct 17 (pp. 1843-1846). ACM.
35. Kanavos A, Perikos I, Vikatos P, Hatzilygeroudis I, Makris C, Tsakalidis A (2014) Modeling retweet diffusion using emotional content. In: IFIP International conference on artificial intelligence applications and innovations. Springer, Berlin Heidelberg, pp 101–110
36. Kim E, Gilbert S, Edwards MJ, Graeff E (2009) Detecting sadness in 140 characters: Sentiment analysis of mourning Michael Jackson on Twitter. Web Ecology. 2009 Aug; 3:1-5.

37. Kim J, Yoo J (2012) Role of sentiment in message propagation: Reply vs. retweet behavior in political communication. In Social Informatics (Social Informatics), 2012 International Conference on 2012 Dec 14 (pp. 131-136). IEEE.

38. Kogan M, Palen L, Anderson KM (2015) Think local, retweet global: retweeting by the geographically-vulnerable during hurricane Sandy. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing 2015 Feb 28 (pp. 981-993). ACM.

39. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer. 2009 Aug 1(8):30-7.

40. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web 2010, ACM, pp 591–600

41. Lee K, Mahmud J, Chen J, Zhou M, Nichols J (2015) Who will retweet this? Detecting strangers from twitter to retweet information. ACM Transactions on Intelligent Systems and Technology (TIST). 2015 May 20; 6(3):31.

42. Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive datasets. Cambridge university press; 2014 Nov 13.

43. Li L, Li A, Hao B, Guan Z, Zhu T (2014) Predicting active users' personality based on micro-blogging behaviors. PloS one. 2014 Jan 22;9(1):e84997.

44. Li LJ, Su H, Fei-Fei L, Xing EP (2010) Object bank: A high-level image representation for scene classification & semantic feature sparsification. In Advances in neural information processing systems 2010 (pp. 1378-1386).

45. Lim KW, Buntine W (2014) Twitter opinion topic model: extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management 2014, ACM, pp. 1319–1328

46. Liu G, Fu Y, Xu T, Xiong H, Chen G (2014) Discovering temporal retweeting patterns for social media marketing campaigns. In: 2014 IEEE International conference on data mining (ICDM) 2014, IEEE, pp 905–910

47. Lu C, Lam W, Zhang Y (2012) Twitter user modeling and tweets recommendation based on Wikipedia concept graph. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence 2012 Jul 15 (pp. 33-38).

48. Luo Z, Osborne M, Tang J, Wang T (2013) Who will retweet me?: finding retweeters in twitter. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval 2013 Jul 28 (pp. 869-872). ACM.

49. Macskassy SA, Michelson M (2011) Why do people retweet? Anti-homophily wins the day!. In Fifth International AAAI Conference on Weblogs and Social Media 2011 Jul 5.

50. Manning CD, Raghavan O, and Schütze H (2008), Introduction to information retrieval 1:496. Cambridge: Cambridge university press, 2008.

51. Mitchell TM. Machine learning. 1997. Burr Ridge, IL: McGraw Hill. 1997;45(37):870-7.

52. Mittal A, Goel A (2011) Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf). pp 15.

53. Mohammad SM, Turney PD (2013) Crowdsourcing a word emotion association lexicon. ComputIntell 29(3):436–65

54. Mohammed M, Khan MB, Bashier EB (2016) Machine learning: algorithms and applications. Crc Press; 2016 Aug 19.

55. Narayanan A, Shi E, Rubinstein BI (2011). Link prediction by de-anonymization: How we won the kaggle social network challenge. In Neural Networks (IJCNN), The 2011 International Joint Conference on 2011 Jul 31 (pp. 1825-1834). IEEE.

56. Naveed N, Gottron T, Kunegis J, Alhadi AC (2011) Bad news travel fast: A content-based analysis of interestingness on twitter. In Proceedings of the 3rd International Web Science Conference 2011 Jun 15, ACM.

57. Nesi P, Pantaleo G, Paoli I, Zaza I. (2018) Assessing the reTweet proneness of tweets: predictive models for retweeting. Multimedia Tools and Applications. 2018 Oct 1; 77(20):26371-96.

58. O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the fourth international conference on weblogs and social media, ICWSM 2010, Washington, DC, USA, 11(122–129):1–2

59. Peng HK, Zhu J, Piao D, Yan R, Zhang Y (2011). Retweet modeling using conditional random fields. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on (pp. 336-343). IEEE.

60. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates. 2001;71(2001):2001.

61. Petrovic S, Osborne M, Lavrenko V (2011) RT to win! Predicting message propagation in Twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media 2011 Jul 5.

62. Pfitzner R, Garas A, Schweitzer F (2012) Emotional divergence influences information spreading in Twitter. In Sixth International AAAI Conference on Weblogs and Social Media 2012 May 20.

63. Piao G, Breslin JG (2018) Learning to Rank Tweets with Author-Based Long Short-Term Memory Networks. In International Conference on Web Engineering 2018 Jun 5 (pp. 288-295). Springer, Cham.

64. Plutchik R (2001) The nature of emotions. Am Sci 89(4):344–350

65. Qiu L, Lin H, Ramsay J, Yang F (2012) You are what you tweet: Personality expression and perception on Twitter. Journal of Research in Personality. 2012 Dec 1;46(6):710-8.

66. Quercia D, Kosinski M, Stillwell D, Crowcroft J (2011) Our twitter profiles, our selves: Predicting personality with twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on 2011 Oct 9 (pp. 180-185). IEEE.

67. Ren F, Wu Y (2013) Predicting user-topic opinions in Twitter with social and topical context. IEEE Trans Affect Comput 4(4):412–24

68. Roberts K, Roach MA, Johnson J, Guthrie J, Harabagiu SM (2012) EmpaTweet: Annotating and Detecting Emotions on Twitter. In LREC 2012 May 21 (Vol. 12, pp. 3806-3813)

69. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents (2004) In Proceedings of the 20th conference on Uncertainty in artificial intelligence 2004 Jul 7 (pp. 487-494). AUAI Press.

70. Sanjari A, Khazraee E (2013) Information diffusion on twitter: the case of the 2013 Iranian presidential election. In Proceedings of the 2014 ACM conference on Web science 2014 Jun 23 (pp. 277-278). ACM.

71. Starbird K, Palen L (2012) (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In: Proceedings of the acm 2012 conference on computer supported cooperative work 2012 Feb 11, ACM, pp 7–16

72. Starbird K, Palen L(2010) Pass It On? Retweeting in mass emergency. In: Proceedings of the 7th international ISCRAM conference seattle 2010 May, vol 1, pp 1–10

73. Stieglitz S, Dang-Xuan L (2012) Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In System Science (HICSS), 2012 45th Hawaii International Conference on 2012 Jan 4 (pp. 3500-3509). IEEE.

74. Sugiyama M (2015). Introduction to statistical machine learning. Morgan Kaufmann; 2015 Oct 31.

75. Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In Social computing (social com), 2010 IEEE second international conference on 2010 Aug 20 (pp. 177-184). IEEE.

76. Sumner C, Byers A, Boochever R, Park GJ (2012) Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In Machine learning and applications (icmla), 2012 11th international conference on 2012 Dec 12 (Vol. 2, pp. 386-393). IEEE.

77. Sun T, Zhang M, Mei Q (2013) Unexpected relevance: An empirical study of serendipity

in retweets. In Seventh International AAAI Conference on Weblogs and Social Media 2013 Jun 28.

78. Teh YW, Jordan MI, Beal MJ, Blei DM (2012) Hierarchical dirichlet processes. Journal of the American Statistical Association, Vol. 101, No. 476 (Dec., 2006), pp. 1566-1581

79. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. Journal of the Association for Information Science and Technology. 2010 Dec 1;61(12):2544-58.

80. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. Fourth Int AAAI Conf Weblogs Soc Media 10(1):178–85

81. Uysal I, Croft WB (2011) User oriented tweet ranking: a filtering approach to microblogs. In Proceedings of the 20th ACM international conference on Information and knowledge management 2011 Oct 24 (pp. 2261-2264). ACM.

82. Vougioukas M, Androutsopoulos I, Paliouras G (2017) Identifying retweetable tweets with a personalized global classifier. arXiv preprint arXiv:1709.06518. 2017 Aug 21.

83. Wang M, Zuo W, Wang Y (2015) A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. Mathematical Problems in Engineering. Volume 2015 (2015), Article ID 936397, http://dx.doi.org/10.1155/2015/936397

84. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. nature 1998 Jun 4; 393(6684):440.

85. Xu Z, Yang Q (2012) Analyzing user retweet behavior on twitter. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) 2012 Aug 26 (pp. 46-50). IEEE Computer Society.

86. Xu Z, Zhang Y, Wu Y, Yang Q (2012) Modeling user posting behavior on social media. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval 2012 Aug 12 (pp. 545-554). ACM.

87. Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z (2010) Understanding retweeting behaviors in social networks. In Proceedings of the 19th ACM international conference on Information and knowledge management 2010 Oct 26 (pp. 1633-1636). ACM.

88. Yarkoni T (2010) Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of research in personality. 2010 Jun 1;44(3):363-73.

89. Zaman TR, Herbrich R, Van Gael J, Stern D (2010) Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips 2010 Dec 10; 104(45):17599-601.

90. Zhang J, Liu B, Tang J, Chen T, Li J (2013) Social influence locality for modeling retweeting behaviors. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2761-2767). AAAI Press [64]

91. Zhang P, Wang X, Li B (2013) On predicting Twitter trend: important factors and models. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2013 (pp. 1427-1429).

92. Zhang Q, Gong Y, Guo Y, Huang X (2015) Retweet behavior prediction using Hierarchical Dirichlet Process. In Twenty-Ninth AAAI Conference on Artificial Intelligence 2015 Feb 9 (pp. 403 - 409).

93. Zhang Q, Gong Y, Wu J, Huang H, Huang X (2016) Retweet prediction with attention-based deep neural network. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management 2016 Oct 24 (pp. 75-84). ACM.

94. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In European Conference on Information Retrieval 2011 Apr 18 (pp. 338-349). Springer, Berlin, Heidelberg.

95. Zhao XW, Guo Y, He Y, Jiang H, Wu Y, Li X (2014) We know what you want to buy: a demographic-based system for product recommendation on microblogs. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014 Aug 24, ACM, pp 1935–1944

96. Zhao Z, Zhan H, Meng L, Xiao J, Yu J, Yang M, Wu F, Cai D (2018) Textually Guided Ranking Network for Attentional Image Retweet Modeling. arXiv preprint arXiv:1810.10226. 2018 Oct 24.