COMBINING TOPOLOGY-BASED & CONTENT-BASED ANALYSIS FOR FOLLOWEE RECOMMENDATION ON TWITTER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYSU YANAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
INFORMATION SYSTEMS

APRIL 2015

Approval of the thesis:

COMBINING TOPOLOGY-BASED & CONTENT-BASED ANALYSIS

FOR FOLLOWEE RECOMMENDATION
ON TWITTER

Submitted by **AYSU YANAR** in partial fulfilment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, Graduate School of **Informatics**

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, **Information Systems**

Assoc. Prof. Dr. Pınar Karagöz _____
Supervisor, **Computer Engineering Department**

Asst. Prof. Dr. Tuğba Taşkaya Temizel _____
Co-advisor, **Information Systems Department**

Examining Committee Members:

Prof. Dr. Yasemin Yardımcı Çetin _____
Informatics Institute, METU

Assoc. Prof. Dr. Pınar Karagöz _____
Computer Engineering Dept., METU

Assoc. Prof. Dr. Aysu Betin Can _____
Informatics Institute, METU

Assoc. Prof. Dr. Halit Oğuztüzün _____
Computer Engineering Dept., METU

Asst. Prof. Dr. Erhan Eren _____
Informatics Institute, METU

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: AYSU YANAR

Signature:

ABSTRACT

COMBINING TOPOLOGY-BASED & CONTENT-BASED ANALYSIS

FOR FOLLOWEE RECOMMENDATION ON TWITTER

Yanar, Aysu

M.S., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Pınar Karagöz

Co-Supervisor: Asst. Prof. Dr. Tuğba Taşkaya Temizel

April 2015, 94 pages

Twitter has become an important social platform for individuals and people share a high number of information about their personal lives, interests and viral news during emergencies. As of 2014, Twitter has 240 million active users and approximately 500 million tweets are shared every day. This information overload in Twitter has become a serious problem due to the growing volume of messages and increasing number of users. Recommender systems help to overcome this challenge.

Finding interesting users and getting useful information from micro-blogging sites has become difficult since the mass of the data contains irrelevant messages, promotions and spam. In this thesis we propose a followee recommender system to overcome this problem. Recommendation in Twitter has been studied by several researchers and promising results have been achieved. In this thesis, we combine topological approaches and content- based analysis within the scope of English and Turkish language to find relevant followees for Twitter users. We propose seven different strategies by using different aspects of Twitter. Personalized recommendations have been generated for 22 active Twitter users. In order to increase effectiveness of recommendations, real Twitter data has been used. The experimental results show that using retweet data gives better recommendations than favorite data and we have achieved 0.79 success rate when we combine the topological features of Twitter.

Keywords: Recommender system, Twitter, Followee Recommendation, Collaborative Filtering, Content Analysis, Topic Mining, Sentiment Analysis

ÖZ

TWITTER İÇİN TOPOLOJİ VE İÇERİK ANALİZİNE DAYALI TAKİPÇİ ÖNERİ SİSTEMİ

Yanar, Aysu

Yüksek Lisans, Bilişim Sistemleri Ana Bilim Dalı

Tez Yöneticisi: Doç. Dr. Pınar Karagöz

Ortak Tez Yöneticisi: Yard. Doç. Dr. Tuğba Taşkaya Temizel

Nisan 2015, 94 sayfa

İnternet kullanımının artmasıyla, insanlar sosyal medya üzerinden gün geçtikte daha çok bilgi paylaşmaya başlamışlardır. Bu paylaşılan veriyle beraber "Aşırı Bilgi Yükleme" problemi ortaya çıkmıştır. Öneri Sistemleri, bu problemin üstesinden gelmek için sosyal medyada sıklıkla kullanılmaktadır. Günümüzde sosyal medyada bilgi paylaşımını sağlayan en önemli kanallardan birisi Twitter'dır. 2014 yılı itibariyle günlük aktif kullanıcı sayısı 240 milyona ve gün içinde atılan tweet sayısı 500 milyona ulaşmaktadır. Bu bilgi trafiği içerisinde ilgi çekici kullanıcılar bulmak ve anlamlı veriyi ayırt etmek oldukça zordur. Twitter üzerinde çeşitli kullanıcı öneri sistemleri daha önce yapılmıştır. Bu tez çalışmasındaki amacımız, kişiler arasında sosyal bağları temel alıp, Türkçe ve İngilizce veriler üzerinde uygulanabilen içerik analiziyle zenginleştirdiğimiz bir kullanıcı öneri sistemi geliştirmektir. Bu çalışmada, Twitter' in farklı topolojik özelliklerini kullanarak yedi farklı strateji geliştirdik. Önerdiğimiz stratejileri test etmek için yaptığımız deneylere 22 aktif Twitter kullanıcısı katıldı. Bu deneylerde her katılımcıya özel olarak, gerçek Twitter bilgisiyle oluştuğumuz öneriler, katılımcıların önerisine sunuldu. Deneylerimizin sonucunda retweet bilgisini kullanarak önerdiğimiz kullanıcıların favorite bilgisinden daha çok tercih edildiğini gördük. Topoloji bilgilerini birleştirerek oluşturduğumuz strateji, 0.79 başarı oranıyla önerilen stratejiler arasından en iyi sonuçları vermiştir.

Anahtar kelimeler: Twitter, Kullanıcı Öneri Sistemi, Kolaboratif Filtreleme, İçerik Bazlı Filtreleme, Güven Verisine Dayalı Sosyal Ağlar, Duygu Analizi

To my beloved husband.

ACKNOWLEDGEMENTS

I would like to take this opportunity to acknowledge and appreciate the efforts of the people who have helped me during my research and documenting this thesis.

Firstly, I would like thank my advisor, Assoc. Prof. Dr. Pınar Karagöz for her support and supervision throughout this research and my co-supervisor Asst. Prof. Dr. Tuğba Taşkaya Temizel for her invaluable criticisms and feedback throughout the study.

My deepest indebtedness is to my family for their love and tolerance. I am grateful to them for their encouragement in every phase of my life. I would like to send all my love and appreciation to my parents Havva Dağlı and Vural Dağlı and my brother Bayram Anıl Dağlı for their endless love, trust and support.

I would like to thank to the members of my thesis examining committee, Prof. Dr. Yasemin Yardımcı, Assoc. Prof. Dr. Aysu Betin Can, Assoc. Prof. Dr. Halit Oğuztüzün and Asst. Prof. Dr. Erhan Eren for their suggestions and constructive criticism.

Very special thanks to my friends Didem, Yücel, İbrahim, Mustafa, Candan, Ergin, Serhan, Burak, Adil, Erinç, Kürşad, Şerife, Çağrı, Ozan, Ömer, Onur, Kübra, Meryem, Vedat, Zeynep and Deniz for their patience and support during the continuous experiments.

Finally, heartfelt thanks to my husband Barış Yanar for his continual encouragement, his tireless assistance and helpful suggestions. I am deeply indebted to him for his endless love, patience and support through all the stages of my study.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

LDA	Latent Dirichlet Allocation
MAP	Mean Average Precision
RT	Retweet
TT	Trending Topic
RS	Recommender System

CHAPTER 1

INTRODUCTION

Recommender systems (RS) have arisen to provide useful and relevant suggestions to the users of various web applications. They are widely used in social networking platforms to find valuable information from large amount of data.

Twitter is one of the most popular social networking platforms in the world. As reported by a recent research in 2014 [35], it has 240 million active users and approximately 500 million tweets are shared by these users every day. This information overload in Twitter has become a serious problem due to the growing volume of messages and increasing number of users. Recommender systems help to overcome this challenge.

In Twitter, people get information from their followees based on their personal interests. In order to get the most beneficial information, active users carefully choose their followees. Searching desirable people is time consuming and not practical in Twitter. In this study, our aim is to help active Twitter users to find interesting people and countervail the information overload problem.

We propose various different strategies in order to find most valuable users by using real Twitter data. We explore two key features of Twitter; the relationships between users and the generated content (tweets). Twitter is a directed social-graph that is composed by followee-follower relationship between users. Several studies in the literature [3, 4, 31, 32, 42] use Twitter's social-graph for recommending new users in Twitter. Previous studies [3, 4, 58, 42] showed that topologically closeness had a positive effect on Twitter users. In addition to followee-follower relationships; retweets and favorites also show the interaction between users [42, 31]. In topology part of our study, we examine these relationships in order to find the best approach for a followee recommender system for Twitter.

Every day, high amount of data are generated from many information sources such as normal users, bloggers, journalists, media institutions etc. Processing these growing data to extract topics becomes significantly important for the recommender systems in order to provide tailored suggestions. To get more personalized

recommendations, we include content analysis into our research for finding more relevant people. In content analysis part of this study, topic mining and sentiment analysis methods are used. In topic mining part, topics are extracted from tweets in order to find topical similarities between users and in sentiment analysis part, sentiment values of tweets are calculated in order to find opinion similarities between users. Although there are several strategies applied on English tweets, for Turkish very few content-based strategies [56, 57] had been applied for followee recommendation in Twitter.

Many methodologies were developed in order to find new followees in Twitter. Some of them used only topology base algorithms [3, 42, 31, 32]. In topology based recommendation strategies, existing links between users were used, such as followee-follower friendship [3], retweet and favorite [42]. Some recommender system applied collaborative filtering approaches such as popularity [55]. In our proposed system, we combine and compare topology based algorithms with content based analysis.

In this thesis work, retweet and favorite data are used separately in our recommendations. In order to understand retweeting and favoriting behaviours in Twitter, a survey is conducted. Our survey results show that people tend use favorite when they like the tweet and find it interesting. Users retweet a tweet when they want to broadcast the information or quote someone else's tweets.

At the end of this study, we evaluate the results of several experiments in order to compare the proposed strategies. Our experiments show that using retweet data gives better recommendations than favorite data. We see a better recommendation performance when we include the topological approaches. Lastly, we do not observe any improvement when we include content analysis together with the topological analysis.

1.1 Contribution of the Thesis

The results of this study contribute to the existing literature as follows;

Firstly, the experiments and the survey results show the behavioural differences between retweeting and favoriting and the effects of these links on followee recommendation on Twitter with using real data. In previous studies [42], favorite and retweet data were used on link predictions with using large data sets however in our experiments, we analyse actual relationships by generating personalized recommendations by using real Twitter data.

Secondly, although some content based strategies are applied on Turkish tweets [46, 56, 57], to the best of our knowledge, this is the first study to combine topical analysis with sentiment analysis on tweets for Turkish language. In our experiments, we compare the topical similarity based strategies and opinion based strategies on followee recommendation in Twitter.

1.2 Outline of the Thesis

The outline of this thesis is as follows:

Chapter 2 – Background Information provides a general overview about recommender systems and Twitter. Definitions of common terms that are used in recommender systems and Twitter are given. Additionally, in content analysis part, LDA and NLP methodologies are explained.

Chapter 3 – Related Work presents the related work on followee recommender systems, topic modelling techniques and sentiment analysis both in English and Turkish. Additionally, in this section, we describe Twitter's current system for topic modelling and recommender systems.

Chapter 4 – Methodology presents our proposed approach in this study. The system design and the modules of our system are explained in detail.

Chapter 5 – Survey states the details of conducted survey. Results are declared and evaluated.

Chapter 6 – Experiment presents the evaluation tools and approaches used in proposed system.

Chapter 7 – Results and Discussion explains the results of the experiments. Applied strategies are compared and evaluated.

Chapter 8 – Conclusion provides a brief summary of the thesis. Additionally, possible improvements are stated.

CHAPTER 2

BACKGROUND INFORMATION

In this chapter, general concepts of the technologies that are employed in this thesis work are provided. In Section 2.1, an overview of technologies for recommender systems is presented. A summary of general information about Twitter is given in Section 2.2. Lastly, about content analysis section, Latent Dirichlet Allocation (LDA) and basic tasks in Natural Language Processing (NLP) are presented in Section 2.3.

2.1 Recommender Systems

A recommender system (RS) is a software tool that provides predictions about a user's potential choices that he/she may be interested in. In this section we discuss the technology behind the recommender systems.

2.1.1 Why Do We Need Recommender Systems?

In our daily life, we are stuck in between choices. What should I wear for work? Which cell phone should I buy? Which movie is the best? Or before buying a product, we feel the need for asking other people's opinions. Recommender systems fill this gap by helping people find the most suitable item for them. Recommender Systems are widely used in web applications in e-commerce. Schafer states that [26] with the help of e-commerce, companies can provide more options to their users. To increase the customer satisfaction and to sell more products, companies have started to use recommender systems [26]. Companies use recommender systems in order to predict their customer's next actions in online transactions. Systems are trained with customer's previous activities. This data is used to understand their customer's preferences and to give suggestions that are more likely to be chosen among the other products/items etc. [26]. For example, at Amazon.com, items are recommended to customers based on other customers' opinions where customers can also rate the user reviews while checking a product. In social media, such as Twitter, users' characteristics are aimed to be analysed for recommendation similar events or users [33, 35]. In addition to this, Facebook has a friend recommender system based on network structure. At YouTube, video recommendations for each user are based on user's latest activities [24]. Google's search engine generates recommendations according to users' location and users' search history [25].

2.1.2 Terms and Concepts in Recommender Systems

According to the Ricci's Recommender Systems Handbook [22], in a recommender system, there are two main data types, "items" and "users. "Item" is used as a general term in recommender system that indicates an object that is recommended to a "User". These items can be movies, news, songs, friends or followees as in our studies which are outputs of a recommender system. At the end of the filtering process, these items are listed for users' considerations. The interaction of items and users is called "Preferences". In addition, users' historical data, i.e. user's previous choices, are used as input in most of the recommender systems. Recommender systems are trained with users' previous preferences. "Rate "is an evaluation result that shows the User's interest on a particular item.

The formal definition of recommendation is as defined below [22]:

Equation 1 shows the utility function equation. Let **G** be the set of all target users. Let us **I** be the union of all items that will be recommended to a user. In order to evaluate the interest of target user **u** to item **i,** the utility function **q** is used:

$$q = GxI \rightarrow L \tag{1}$$

Equation 2 shows the formal definition of recommendation problem. Let **q** be the utility function that shows the benefit of the item **i** to user **g** and **L** is a ordered item set. Each element in **L** is a possible item **i** to be recommended to target user **g**.

In other words, in every recommender system most useful item $i' \in \mathbf{I}$ is recommended for each user $\mathbf{g} \in \mathbf{G}$, to maximize user satisfaction.

$$\forall g \in G, i \in I, i'_g = argmax\, q(g, i) \tag{2}$$

Each user has a profile that is used as an "input" in recommender systems. These profile data are commonly categorized as follows:

Ratings:

Ratings/Votes show a user's opinions on a particular item. Ratings could be in four types [28],

Numerical – scalar numerical ratings, such as 1 to 5

Ordinal –scalar sequenced ratings such as strongly agree, agree, neutral, disagree and strongly disagree

Binary – allows only two options such as agree /disagree, 0-1, like/dislike

Unary – if a user prefers or purchases the particular item, rating is evaluated as positive.

Demographic data:

Demographic data indicates the quantifiable statistics of a profile. Age, city, education, gender and nationality are considered as demographic properties.

Content Data:

Content data is pertaining to texts that are related to an item or user. Content data could be about rating a product or could be personal data which is shared in social media. Researchers use this social media profile data for understanding tendencies.

The output of a recommender system could be a *prediction* or a *recommendation*. The difference between these two concepts is that the result of the *prediction* is the user's expected choice for a particular item. On the other hand, a recommendation is a list of items that shows the possibilities that users might prefer.

Neighbourhood/Topology:

Neighbourhood is a virtual space/graph formed by user relationships. The relationships between the users, could give information about user's preferences. For example, Twitter's user recommender system is based on a directed and weighted graph that shows the similarities between users [32].

2.1.3 Types of Recommender Systems

Ricci et al. [22] classified recommender systems according to their prediction technique. Six types of personalized recommender systems were introduced based on data types and usage as, collaborative recommender systems, content-based recommender systems, knowledge-based recommender systems, demographic recommender systems, community-based recommender systems and hybrid recommender systems.

2.1.3.1 Collaborative Recommender Systems

The aim in collaborative recommender (CR) systems is to find the most valuable option for a specific user by collecting the data from other users' previous preferences or interactions. Collaborative Filtering (CF) technique is commonly applied in collaborative recommender systems. The main objective of CF is finding the most acceptable alternative among the user-item pair list. In CR systems, firstly users are modelled according to their previous choices. Secondly, users are grouped depending on similarity of taste to other users. Lastly the system suggests items that have been rated/bought by similar users. More generally, collaborative systems are known as "People's Choice". CF is most widely used technique among other recommender systems [22]. In Amazon.com, while a user is checking an item, some

recommended products are shown at the bottom of the page. These products are those that were purchased together with the current product previously by other users. YouTube recommends a few similar videos based on previous personal activities such as adding favorites, giving ratings etc. and other user's previous choices that listened the same songs formerly.

2.1.3.2 Content-Based Recommender Systems

Content-Based Recommender Systems (CBRS) are based on keywords in the text data. These data are collected from item descriptions, user's previous preferences or user profiles. CBRS recommends items regarding the user's previous choices or ratings.

Information retrieval and information filtering concepts are commonly used in CBRS. The collected data could be combination of relevant and non-relevant noise data. Information filtering operation is used for weeding out these noise and unessential data from data chunk. After information filtering process, information retrieval is used for indexing and finding the most reliable information. In CBRS, users and items are represented by keyword vectors. The main goal is finding the most similar pair of user and item vectors that based on the collected data from user's previous preferences.

2.1.3.3 Knowledge-Based Recommender Systems

In Knowledge-Based Recommender Systems (KBRS), recommendations are generated in order to fulfil user needs, based on gaining knowledge from the data of users and items. According to Burke's study [37], KBRS are commonly used when collaborative filtering and content-based filtering cannot be performed. Both in collaborative filtering (CF) and content-based filtering (CBF), user ratings are collected to train the system. Insufficient amounts of data could be misleading in training process, thus CF and CBF need large amount of data to train the system. KBRS can be applied in data shortage cases like cold-start problems.

2.1.3.4 Demographic Recommender Systems

Demographic Recommender Systems (DRS) are based on demographic characteristics of human beings such as age, race and gender. Considering demographic properties has significant effects on accuracy of RS. The advantage of the DRS is that there is no need for user rating data as in the Collaborative Filtering or in the Content Based Filtering Systems [38]. DRS are used in e-commerce, for instance, recommending books based on gender or showing search results according to location.

2.1.3.5 Utility-Based Recommender Systems

Utility-Based (UB) Recommender Systems are based on a quantitative approach that indicates the calculation of utility values. This utility value is calculated by including all properties of items for each user. Besides tangible properties, intangible properties like product or user reliability could be included in calculation of utility value. UB recommender systems do not generate long term recommendation but rather they recommend according to user's current needs and available options at that time. The advantage of this approach is the ability to include all considerations of a user into the system and the utility values are specific for each user [38].

2.1.3.6 Hybrid Recommender Systems

Hybrid recommender systems are combination of two or more recommender systems that are mentioned above. In order to gain better performance, integrated methodologies are used in recommender systems. Generally, collaborative filtering techniques are combined with other techniques to minimize the errors. According to Burke [38], there are seven types of hybrid RS. These are weighted, mixed, switching, feature, combination, cascade, feature augmentation and meta-level.

1. The *weighted hybrid recommender system* is based on item ratings that are calculated from all recommender systems in the system. For instance, weighted method is useful for adjusting the scores of recommender systems in linear calculations.

2. The *mixed hybrid recommender system* is based on generating recommendations from all different kinds of techniques at the same time.

3. The *switching hybrid recommender system* is adaptable in changing circumstances. The applied recommendation technique is switched in parallel to the transforming conditions.

4. The *feature combination hybrid recommender system* is based on combining different capabilities in different RS in a single RS. Our approach in this study is mainly based on feature combination.

5. The *cascade* is a phase based recommender system. After applying first recommendation, the second one is applied to enhance the recommendation from the possible suggestions.

6. In *feature augment*ation hybrid recommender systems, the result of the first recommendation is used as an input of second recommender system.

7. *Meta-level hybrid recommender system* is similar to feature augmentation hybrid RS. Both use the results of the first recommender system. The difference in meta-level hybrid RS is training system for generating recommendations for second model. By means of this the entire model development depends on user inputs.

2.2 Twitter

2.2.1 What is Twitter?

Twitter is an online social networking service, which was created in October, 2006 by Jack Dorsey, Even Williams and Biz Stone. According to [17], Twitter can be used for many different purposes. One purpose is providing a micro blogging service for sharing details of a person's life. Moreover, Twitter can be used as a marketing tool for public relations as in many politicians and celebrities use for interacting with their audience. On many occasions, Twitter can also be used as a social messaging service that enables interactions among users. People can communicate with their friends and family and share details of their lives. Lastly, Twitter is an information platform on which users can get news via broadcasting agents' or journalists' accounts fast and easily rather than watching television or reading the newspaper. Moreover, information can spread very quickly though Twitter.

2.2.2 Twitter Glossary

In this section, we aim to give basic information about this special lingo of tweeting.

- *Tweet* is mainly a short message which is an expression of a moment or idea that can be posted on Twitter and should be less than 140-character. The beauty and the challenge of Twitter is 140 characters limitation. This means that when someone wants to say something on Twitter, it has to be less than 140 characters.
- *Retweet* is sharing a tweet with third parties. According to Bongwon's study [18] retweeting is the base mechanism for information diffusion. Generally, retweeted tweet means some interesting information that worth to be shared.
- Favoriting a Tweet can let the original poster know that you liked their Tweet, or you can save the Tweet for later[1].
- *Following* someone means that choosing/asking an individual to receive status updates on current timeline.
- A *follower* is a person who follows a user to get updates on their own timeline.
- A *followee* is a person who has chosen another Twitter user from whom to receive one's updates in their timeline.
- *Unfollowing* is choosing not to receive one's tweets on timeline.
- *Mention* is any Twitter update that contains "@username" anywhere in the body of the Tweet [19].
- *Hashtag* is used to tag certain events or contexts. Hashtag can be a word or combination of words, is used with prefix symbol "#"and a keyword. Hashtags mostly show the topic of the tweet [20]. Hashtags can be used for detecting trending topics and it can be used for coordinating distributed discussions or spread information through large groups that are not connected.

[1] https://support.twitter.com/articles/20169874-favoriting-a-tweet

- *Trending Topic (TT)*: if a particular term is significantly mentioned among other topics then it becomes a trending topic. TTs can be phrases, words or hashtags. Twitter collects these terms and shows the list of most popular top ten trending topic on the main page of Twitter based on user's location [21].

2.2.3 Twitter API

Twitter API allows users to obtain data from twitter servers. Twitter provides two APIs for collecting data REST API[2] and Streaming API[3]. In Streaming API, Twitter provides continuous access to users who want to gather data from Twitter's global stream of Tweet data. Rest API provides restricted access and controls the data collection amount with rate limits. In spite of the limitations, a user can demand more specific data in Rest APIs such as the tweets from a particular time or the profile data from a certain user.

2.3 Content Analysis

With the exponential growth in the usage of the Internet, information generation has increased. Therefore information retrieval from large text collections has become more important. Machine learning is one of the commonly used techniques to extract information from big data. There are two types of machine learning techniques; supervised and unsupervised. In supervised machine learning, a labelled training data is used for inferring function maps. In the training data, each data point is associated with a label. In unsupervised learning, data are not labelled and the goal is to find hidden patterns and structure from unlabelled data.

Obtaining labelled training data from unstructured text collections is costly and difficult in supervised machine learning. To overcome this challenge, unsupervised machine learning techniques are used for analysing unlabelled big data.

LDA is one of the most powerful unsupervised machines learning technique that is used for extracting topics from large data sets [10]. There are also efforts to develop new models by extending LDA for different purposes.

Sentiment analysis over Twitter helps to analyse people feelings towards topics. In order to extracting sentiments from tweets, NLP is used for differentiate languages processing words and for morphological analysis of tokens in the text. In this study, Zemberek library [12] is used for NLP of Turkish texts.

The detailed information about LDA, NLP and the derived methodologies are given in this section.

[2] https://dev.twitter.com/rest/public

[3] https://dev.twitter.com/streaming/overview

2.3.1 Latent Dirichlet Allocation (LDA)

Blei et al. [10] proposed an unsupervised, non-parameterized and generative topic modelling called Latent Dirichlet Allocation (LDA). LDA is widely used to extract topical key phrases from large datasets. Typically, list of topical terms is listed as key phrases which indicate the topics of a document. LDA is based on "bag of word" approach where every document is shown as a vector of words. In LDA based approach each document is composed of probability distribution of various topics and these topics are composed of probability distribution of words. A topic is defined as a distribution over word corpus.

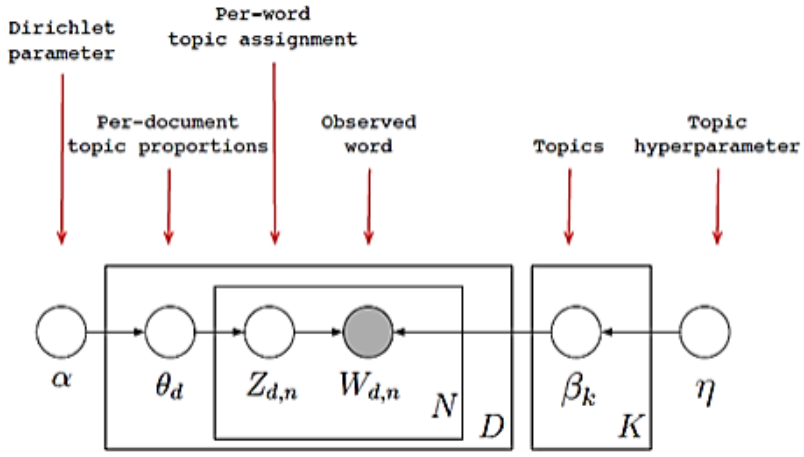


Figure 1: Plate notation for LDA

Figure 1 shows the plate notation of LDA. Formally LDA can be defined as follows: Let there be *K* topics , *d* is a document and *w* is a corpus of words from corpus of documents *D* and each word in a document ($w_{d,n}$) is an element of word corpus where *N* plate shows the collection of words in Document *D*. α is a Dirichlet hyperparameter for multinomial per-document-topic distribution (θ_d) and η is topic hyperparameter for multinomial word-topic distribution (β_k) over vocabulary *N*. $z_{d,n}$ indicates the assignment of a topic for *nth* word in a document *d*. w_d refers to words of the document *d*.

The pseudo code of LDA from [10] is as follows:

//topic:

1. For each topic z= 1,…, K

Choose mixture components $\beta_k \sim Dirichlet(\eta)$

//document:

2. For each document d = 1,...,D

Choose mixture distribution $\sim Dirichlet(\alpha)$

//word:

For each word $w_{d,n}$ in a document n= 1,…,N

1. Choose a topic $z_{d,n} \sim Multinominal(\theta_d)$

2. Choose term for word $w_{d,n} \sim Multinominal(\beta_{z_{d,n}})$.

2.3.2 Author- Based LDA

Rosen-Zvi et al. [11] proposed an author-topic model for topic discovery from large texts. Author-Based LDA is an extended version of LDA that adapts the authorship information to LDA [11]. Figure 2 shows the plate notation of Author-Based LDA

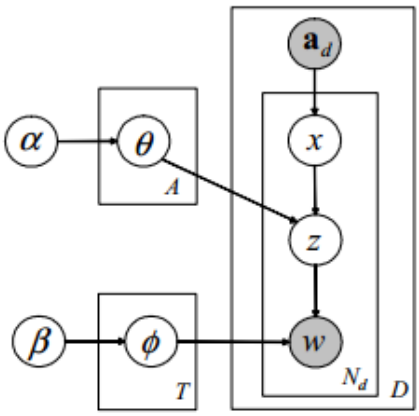


Figure 2: Plate notation for Author-Topic Model LDA

According to Rosen's study [11], A is the corpus of authors and a_d shows the sets of authors of document d. Each author in a_d chooses topics from distribution over topics θ and generates a document d that is a collection of words w_d. Each word w in a document d is associated with a topic. A topic is selected from distribution over topics specific to an author and each author x is associated with a distribution over topics θ. The pseudo code of author-topic model from [11] is as follows:

1. For each author a= 1,…,A choose W dimensional $\theta_a \sim Dirichlet(\alpha)$

   For each topic t= 1,…,T choose W dimensional $\emptyset_t \sim Dirichlet(\beta)$

2. For each document d = 1,…, N_d

   From a group of authors a_d,

3. For each word $w_{d,n}$ in a document n= 1,…,N

   1. Choose a random author $x_{d,n} \sim Uniform(a_d)$
   2. Choose a random topic $z_{d,n} \sim Discrete(\theta_{x_{d,n}})$
   3. Choose a random word $w_{d,n} \sim Discrete(\beta_{z_{d,n}})$

2.3.6 Natural Language Processing

Natural Language Processing (NLP) is a research area of computer science which gives the ability of processing and understanding human generated languages. The aim in NLP process the interaction between computers and human, understands the linguistic details of the languages and developing software based automated solutions for language processing

NLP in Turkish is quite challenging because of the complex morphology of Turkish [39].Turkish is an agglutinative language where a word can be decomposed to its root and derivational suffixes. Every suffix can be used for indicating tenses, personality, negation, etc. For instance, the translation of "Malezyalılaştıramadıklarımızdanmışçasına" is "as if you were one of those whom we could not make resemble the Malaysian people".

Zemberek [12] is an open source JAVA library, provides a comprehensive NLP framework for Turkish language. It provides spell checking, morphological parsing, stemming, natural language understanding, spell checking and Part-of-Speech tagging for Turkish language. In Figure 3, the POS stemming suggestions are shown for "güvenilirlik" expression. The result shows the type of the root of the word and the usage of each suffix.

```
güvenilirlik:
[ Kok:güvenilir, Tip:ISIM | Ekler:ISIM_KOK, ISIM_BULUNMA_LIK]
[ Kok:güvenilir, Tip:ISIM | Ekler:ISIM_KOK, ISIM_DURUM_LIK]
[ Kok:güven, Tip:FIIL | Ekler:FIIL_KOK, FIIL_EDILGEN_IL, FIIL_GENISZAMAN_IR, ISIM_DURUM_LIK]
```

Figure 3: Zemberek Suggestions for the word "güvenilirlik"

CHAPTER 3

RELATED WORK

In this chapter, the related works are summarized under two subsections. The recommender systems studies conducted on Twitter data are given in Section 3.1. The studies on topic modelling and sentiment analysis are presented in Section 3.2.

3.1 Recommender Systems on Twitter

Several studies have been conducted for recommending new followees for Twitter users. In this section, we present summary of related studies on followee recommendation for Twitter.

Armentano et al. [3] used a topology based algorithm for followee recommendation for Twitter. The authors explored the graph of connections that originated from the target user. They made a list of candidate users that they ranked using different weighting features such as popularity (#followers/#followees), the number of the occurrences of a given user in a list of candidate users and number of common friends. The approach that was described in this paper is similar to ours as to making candidate list and ranking methodology process. Differently, in addition to followees of followees' data, we include retweets of retweets and favorites of favorites data in our study.

In Garcia's study [5], the authors proposed a recommender system based on the relation between popularity and activity. They calculated a popularity threshold (#followees/#followers) and activity range (range of total number of posts) for filtering the recommendations. In their system, if the target user's followees' popularity ratios were higher than the threshold, they recommended popular users to the user. If the target user's followees had high activity ratio, they recommended active users to the user. They observed that they had better results when they used popularity and activity together. In this paper, they included collaborative approaches in followee recommendation in Twitter. However, we did not use any collaborative approaches in our study.

Golder et al. [55] used four methods for identifying whom users might want to follow. These methods were 'reciprocity', 'shared interests', 'shared audience' and 'filtered people'. In reciprocity method they assumed that if someone started to follow a user, this user would follow back his or her followers. In shared interests and shared audience methods, they concentrated on homophily between users, which stated that people tended to follow someone who was like-minded or similar to the others. In filtered people method, they filter users whose tweets were retweeted by the followees of this user. The paper also stated that a user might be interested in followees of followees because they might also share the same interest. In our study, we use followees of followee data. Moreover, we use retweets of retweets and favorites of favorites data to generate recommendations.

Tavakolifard et al. [42] proposed a recommender system that focused on hidden relationships between users. The authors stated that users did not interact with all of their followees hence they extracted hidden relationships between Twitter users in their study. They concentrated on four Twitter specific relationship types (followee, follower, retweet and favorite) that indicated trust relationship between users. Firstly, they filtered the social relationships to identify the stronger relationship ties. They mainly concentrated on retweet and favorite behaviour in Twitter. Secondly, they applied four trust propagation methods to extend user's hidden network (web-of-trust). The applied algorithms were simple-transitivity, weighted-transitivity, golbeck-transitivity and structural-similarity. In "weighted-transitivity", they assigned a value to the each relationship link where this value showed the number of other users that connected them in a transitive path A → B → C. In "golbeck-transitivity", they ranked their links according to their trust ability. They assumed that when User B as a referral who sent her opinion of trust about User C to User A , it was important to consider User A's trust in User B and User B's trust in User C. In "structural similarity" SimRank was used in order to find similarities between users. Their results showed that "simple transivity" gave a better coverage than "weighted-transitivity" and "golbeck-transitivity". They found that structural similarity is a better propagation method on a web-of-trust generated by a user's retweet behaviour. The approach considered in this paper is similar to our topological approaches. We use "simple- transitivity" as a base line for our work. Since the graphs that we construct in our study have a maximum path length of two, we do not use SimRank [54]. In our research, we do not only use topological strategies for recommending new users but also use content based strategies to rank the candidate list.

Hutto et al. [4] have explored the relative effects of social behaviour, message content, and network structure on follow behaviour. They have come up with some deductions. They explored that social behavioural choices those may affect network growth, such as, people had the tendency to trust users who filled the profile content in Twitter and their studies have shown that topologically closeness has a positive effect on Twitter users. In our study, we use topologically closeness also as a baseline of our recommendations.

3.1.1 Who To Follow Service in Twitter

Based on [31, 32, 33, 36], Twitter has used machine learning based recommender system. They have combined topological graph based analysis [32] with collaborative analysis [32, 33] in large scale networks [36]. Twitter has introduced who to follow (WTF) service as a followee recommender system [31]. WTF service has relied on personalized Pagerank algorithm which runs on Twitter social graph. While generating this social graph, they have concentrated on two types of relationships which are "similar to" and "interested in". With regards to the shared common interests between users, these two users could be presumed as similar. On the other hand, for example if a user is interested in Zeki Muren, we cannot say that Zeki Muren is similar to our user. WTF Service builds up a Twitter graph that is formed by users connected with edges that shows the followee- follower relationship. They mention that the connections in the graph represent the interest graph, instead of a social graph. The interest group constitutes a "circle of trust", which is developed with user centred random walks, also known as personalized PageRank.

Recently, Twitter has introduced "Similar to" framework to provide better recommendations [32, 33, 36]. In [33], they gave details of the "Similar-to" framework that focuses on discovering top similar users for each type of user in Twitter. As shown in Figure 4, Similar-to framework has three parts, the first part is candidate generation, the second one is model learning and the third one is Regression.

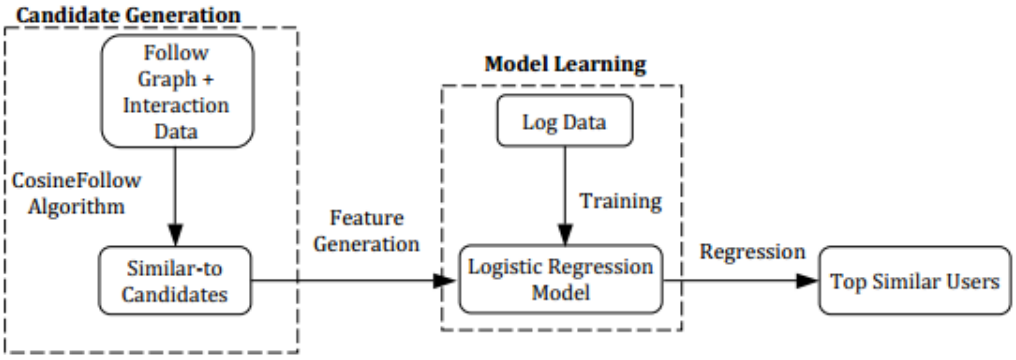


Figure 4 : Twitter Similar-to Framework

Candidate Generation phase was explained in more detail in Kamath's [32] study. In [32], they developed a directed and weighted graph (RealGraph) that showed interactions and relationships between users and possibility of oncoming interactions. Kamath used multidirectional relationship data as an input for RealGraph system. This data contained users' retweets, favorite tweets, email addresses and historical data including users' clicks, viewed contents and viewed retweets. According to this data, they assigned a weight for each connection that indicates the strength of the relationship between users. Furthermore, they included time sensitive data such as when and how often these interactions happen.

According to Lin's study [36], in model generation stage, Twitter used Hadoop-Pig based analytics platform for model training process that is known as streaming training. In this study [36], they concentrated on end-to-end machine learning workflows and integration of these methods to large scale platforms. Stream training platform is used to train model daily for updating the RealGraph edge values. At the last phase, to calculate similarity between nodes, they ran cosine similarity on RealGraph. They found the most similar users according to the score gained from cosine-similarity. To sum up, these proposed works [31, 32, 33, 36] shows the Twitter's current recommender system. They have ranked the relationships between users according to several methodologies based on several features users' retweets, favorite tweets, email address and historical data including users' clicks, viewed contents and viewed retweets. Hence we are not able to use some of these data due to the Twitter API limitations.

3.2 Content Analysis

3.2.1 Topic Modelling with Twitter Data

Despite the 140-character limitation, tweets can have strong meanings. Finding topics from noisy, short and ambiguous Twitter data is quite challenging [35]. Several studies (LDA [44], AT [14], LLDA [45], and Twitter-LDA [13]) have been carried out to extract the correct meaning from tweets. Both Zhao et al. [13] and Hong et al. [14] reformed LDA for Twitter text. In Hong's study [14]; author- topic model was adapted to Twitter texts. In [14] it was assumed that each author had one document and this document contained all tweets from the same user. Based on this assumption, each document could have more than one topic.

3.2.1.1 Twitter-LDA

In opposition to [14] in Twitter-LDA, the assumption is that every status can have only one topic in Twitter due to the character limitation. Each word in a document can belong to a topic or it can be considered as a background word (common word). High-frequency words are assigned as a background word to get more significant topics. Figure 5 shows the Twitter-LDA topic generation methodology from tweets. Formally, β, α, γ are Dirichlet distribution parameters, T is corpus of topics. Let $\emptyset^t$ be the word distribution over topic t and $\emptyset^\beta$ be the background word distribution. Each word in a tweet is decided whether it is background or topical word according to $\emptyset^\beta$ and $\emptyset^t$. Each user chooses topics from distribution over topics $\emptyset^t$ and generates a tweet t that is a collection of words w_i. Let θ^u show the topic distribution of user u. A topic is selected from the distribution over topics specific to a user and each user u is associated with a distribution over topics θ^u. The pseudo code of Twitter-LDA model from [13] is as follows:

1. Draw $\emptyset^{\beta} \sim Dirichlet(\beta)$, $\pi \sim Dirichlet(\gamma)$

2. For each topic $t = 1,\ldots,T$

   (a) Draw $\emptyset^{t} \sim Dirichlet(\beta)$,

3. For each user $u = 1,\ldots,U$

   (a) Draw $\theta^{u} \sim Dirichlet(\alpha)$

   (b) For each tweet $s = 1,\ldots, N_u$

   i. Draw $z_{u,s} \sim Multinominal(\theta^{u})$
   ii. For each word in a tweet $i = 1,\ldots, N_{u,s}$
      A. Draw $y_{u,s,i} \sim Bernoulli(\theta^{u})$
      B. If $y_{u,s,i} = 0$ draw $w_{u,s,i} \sim Multinominal(\emptyset^{\beta})$ and if $y_{u,s,i} = 1$ draw $w_{u,s,i} \sim Multinominal(\emptyset^{z_{u,s}})$

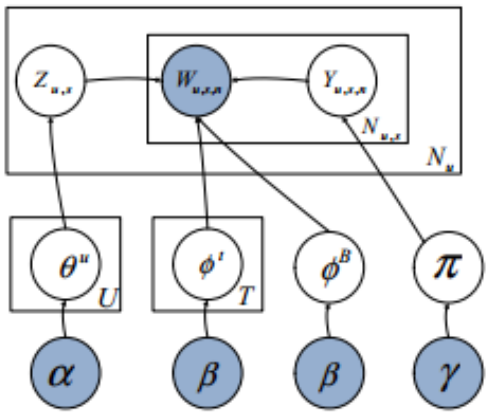


Figure 5: Plate Notation for Twitter-LDA

3.2.2 Topic Modelling System in Twitter

Yang et al. [35] proposed a supervised text classification system that inferred users' interests by processing real time data in Twitter production environment. Their system had two phases; elimination of the non-topical content [34] and assigning topical tweets to predefined topic list [35]. They comprised a predefined ontology that included 300+ topics including hierarchical relationships by referring to the existing ontologies [35]. To achieve high accuracy in topic modelling, a sorting system [34] was proposed to weed out the non-topical, noise data from data chunk.

In Ramnath study [34], they focused on eliminating the non-topical, personal or noisy content from Twitter real time data. Their elimination system, called "Chatter", was based on LDA. Firstly, they applied LDA to get 500 topics from Twitter data.

Secondly, they picked non-topical topics among these topics. At the end they picked 70 personal content topics that had the threshold value higher than 0.70.

Their large scale topic modelling system was used to obtain non-personal and topic concentrated content to assign topics to tweets [35]. For gaining valuable content, they used URLs, user data and hashtags. For instance they chose some specific users who mostly talked about single topic (e.g. @ntvspor) and they used entity level content such as #hashtags that was assumed as user generated topics. Moreover they took advantage of URL containing tweets. It was assumed that the text of URL could give some information about topic (like itunes.apple.com/tr/app/ntv-spo...). They used this non-personal valuable context data to train and integrate their topic modelling system and after that they created relations between the topics and tweets. These studies were helpful in understanding the eliminating chatter content in Twitter. We use some of their approaches in our study such as eliminating non-topical topics; however, they use pre-defined topical ontology that included 300+ topics. Since the topics in Twitter are unsteady and changeable according to location, age, gender etc., it is not possible to get minor topical similarities between users.

3.2.3 Studies of Turkish Texts

In our research, two of our proposed strategies are based on content-based analysis. These proposed content based strategies are applied on both Turkish and English texts. Since Turkish is an agglutinative language, applying content-based algorithms are more challenging than English.

There are several strategies applied for topic mining in Turkish tweets [56, 57]. Gemici et al. [57] applied LDA on Turkish tweets. While collecting data they used twitter4j [49] and Zemberek [12] for stemming. The approach described in this paper is similar to ours as to finding topics from Turkish tweets. We also use twitter4j and Zemberek in topical analysis part of our study. However, instead of LDA, we use Twitter-LDA, which is an enhanced version over LDA for Twitter text.

Celebi et al. [56] proposed a followee recommender system for Twitter that was based on finding behavioural similarities by using tweets and retweets of users. They extracted topics from tweets by using TF-IDF schema. Their system used five different relevancy scores; Feedness, Socialness, Retweeted, Hashtag Usage and Term Variation. Their recommender system was only based on content-based algorithms. In our study, we use not only content based similarities but also topological closeness. To the best of our knowledge, this is the first study to combine topical analysis with sentiment analysis on tweets for Turkish language. In our experiments, we compare the topical similarity based strategies and opinion based strategies on followee recommendation in Twitter.

3.2.4 Sentiment Analysis

Sentiment analysis, also called opinion mining, is used for understanding people's opinions, feelings and thoughts. Opinion mining from user generated content such as tweets, news, and blogs has vital importance on businesses, politics and marketing etc.

“Other people’s opinions” has significant effect on decision-making process. Since the subjectivity is used commonly in social media such as Twitter, many studies have been conducted to extract people’s opinions from large Twitter data sets. Hutto's study [4] showed that the content had significant impact on gaining followees. According to their research [4] the informational content has positive effects on gaining followees; whereas negative sentimental tweets have negative effect.

In Kouloumpis's research [9], existing sentiment analysis methods were tested. In their approach, hashtags were used to detect topics and emoticons were used to detect emotions. Five experiments were applied; N-grams, n-grams with lexicon, n-grams with part-of-speech (POS), and n-grams with lexicon with Twitter features such as emoticons and finally combination of all features. They have used two datasets, HASH and EMOT, which were trained with hashtags and emoticons. They found that n- grams with lexicon features and Twitter features on HASH data set had a better performance than POS and they also discovered that lexicons also had a connection with sentiments but emoticons, abbreviations and intensifiers had better performance than the lexicons.

In Pak's point of view [6], the emoticon in a sentence portrayed the sentiment of the entire sentence. They prepared one positive, one negative and one objective data set that were tagged accordingly. It was aimed to classify data by using methods such as Naive Bayes, SVM and CRF. Their experiments showed that Naive Bayes classifier produced the best results.

The number of sentiment analysis studies in Turkish is limited. In Erogul's study [15], supervised machine learning (ML) techniques were used for sentiment analysis in Turkish. He applied SVM, n-grams, POS (Part-of-speech) tagging and combinations on a labelled movie review dataset for his work and 85%of accuracy was achieved. Kaya [46] applied supervised machine learning techniques to analyse sentiments of political news. Transfer Learning was used to improve the system with transferring useful knowledge from Twitter. Naïve Bayes (NB), Maximum Entropy (ME), SVM and the character based N-Gram Language Model were applied to extract sentiments from political columns. According to the experiments without Transfer Learning, Maximum Entropy and the character based N-Gram Language Model produced better results than NB and SVM. By including Transfer Learning approach, accuracy of results reached 90% for NB, SVM, ME.

Recently, in Turkmenoglu's study [47], lexicon based and Machine Learning techniques have been compared. They have used Twitter datasets for sentiment analysis and movie review datasets for comparing the previously used algorithms. Zemberek [12] has been used for deasciifiying operations. They have applied two-level description of Turkish morphological parser and morphological disambiguator in pre-processing phase. The prediction accuracy in their study is 85% in Twitter dataset and 89% in movie dataset by using SVM.

In order to calculate sentiment values from tweets, SentiStrength was used in this study. Thelwall et al. [47] introduced SentiStrength as a sentiment detection framework for English texts. SentiStrength is a lexicon based sentiment analysis (opinion mining) tool for short social texts. A. Gural et al. [16] developed a lexicon-based framework that used SentiStrength lexicon for Turkish language. In SentiStrength framework, sentiment results are categorized in three way; binary, trinity and single score. Binary results are positive, negative values and trinary

results include positive, negative, neutral values. Lastly, single score results shows the degree of sentiment of a sentence.

The corpus of SentiStrength is generated by five lists; sentiment word list, booster word list, idiom list, negating word list and emoticon list. Sentimental values for lexicon words and symbols are given to calculate the sentiment values for a sentence. SentiStrength uses emotion lookup table that includes sentiment scores of words. The words in this list are scored from +1 (not positive) to +5 (highly positive) for positivity and from -1 (not negative) to -5 (highly negative) for negativity. The sentiment score scale in SentiStrength is shown in Figure 6. Booster words are listed under booster word list to specify the weakness or strengthen the sentiment of the word. Phases are listed under idiom list, which has its own word strengths. The negation word list includes the negation words. Lastly emoticons are listed under emoticon list. Table 1, shows the samples from the lists.

Table 1: Sample Words from SentiStrength Corpus

Word	**English**	**Listed under**	**Value**
dostluk	friendship	EmotionLookupTable	2
nefret	hate	EmotionLookupTable	-4
(-:	(-:	EmoticonLookupTable	+1
değil	not	NegatingWordList	
abuk subuk	incoherent	IdiomLookupTable	-2
en	most	BoosterWordList	1

Table 2 shows the sentiment score from SentiStrength for “Mevcut politik havadan nefret ediyorum”. In analysing part, sentiment verbs are extracted. The meaning of “nefret” word in English is “hate” has been scored as -4 according to emotion lookup table. Since the negativity is higher than positivity in the sentence, binary prediction is scored as -1, which indicates that the sentiment of sentence is negative. As a result, the accuracy of Gural’s unsupervised study [16] is 79% in Turkish movie dataset and 75% in Twitter dataset.

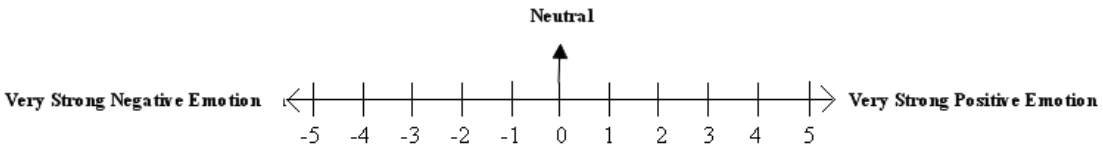


Figure 6: SentiStrength Sentiment Score Scale

Table 2: Sample of SentiStrength Score for Turkish Sentence

Sentence	Analysing	Positive Score	Negative Score	Binary Prediction
Mevcut politik havadan nefret ediyorum. (I hate the current political climate)	Mevcut politik havadan nefret [-4] ediyorum.	1	-4	-1

3.3 Summary

In this section, we have summarized some of the related studies. In the first section, we have reviewed some studies about identifying new followees in Twitter and we have compared them with our study. Studies [3, 4, 55, 42] showed that Twitter users tend to follow other users those are topologically close. Tavakolifard et al. [42] focused on extracting hidden relationships between users like retweets and favorites. In the second section, we have mentioned some studies regarding content based analysis. Some of these presented studies such as SentiStrength and Twitter-LDA are used in this thesis work. In our research, we do not only use topological closeness for recommending new users but also combine topological approaches with content based algorithms for both Turkish and English tweets in order to rank the candidate list.

CHAPTER 4

SURVEY STUDY

This survey is implemented in order to assess a comprehensive overview of favoriting and retweeting behaviour in Twitter. This chapter describes the overall design of survey, the survey questions and evaluation of survey results. Firstly, the design of survey is explained in detail. Secondly the results are presented. The survey items are given in Appendix A.

4.1 Survey Design

At the beginning of this study, a survey is conducted in order to investigate the retweeting and favorited behaviour in Twitter. An online survey is created and hosted by Survey Monkey. The link of survey is distributed though the researcher's Twitter account. The survey was accessible between 1st of March 2015 and 16th of March 2015. There were 73 user accesses to the survey and 8 surveys were removed since they include no data. After cleaning process, 65 surveys were analysed which has 88% effective response rate.

In this survey, 9 items were created based on Twitter usage behaviour. Initially, the survey starts with 6 general Twitter usage questions, including how long and how often the participant has used Twitter, why they use Twitter, how many followers and followees they have. After answering initial questions, in the 7th question participants were asked whether the "favourite" behaviour is similar to "like" behaviour in Facebook or not. In order to analyse the favorite behaviour in Twitter, they were asked 10 favorite related queries in question 8. Lastly, to analyse the retweet behaviours, we asked the participants to rate 10 different possibilities in question 9. .

4.2 Survey Data Analysis

4.2.1 General Information

In the first question, participants were asked their Twitter user name and due to the privacy concerns of some users, this field remained as a non- mandatory field. Among 65 participants, 38 of our participants provided input for this area (58%).

In the second and third question, we asked our participants' range of followee and follower count. Followee and follower counts were measured on a 6-point Likert-type Scale. The scale used was "less than 10 ", "between 10 and 100", "between 1000 and 10000", "between 10000 and 100000" and "more than 100000".

According to Figure 7, all of our participants follow more than 10 users and less than 10000 users. The majority of followee counts are between 100 and 1000. Figure 8 shows that among 65 participants, only one participant has less than 10 followers and 38 of the participants (58%) have followers between 100 and 1000.

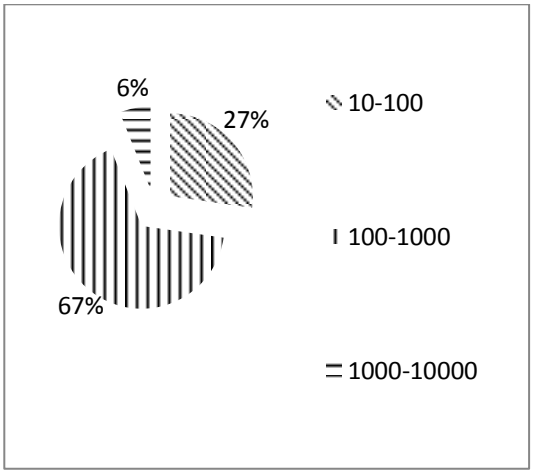


Figure 7: Number of Followees

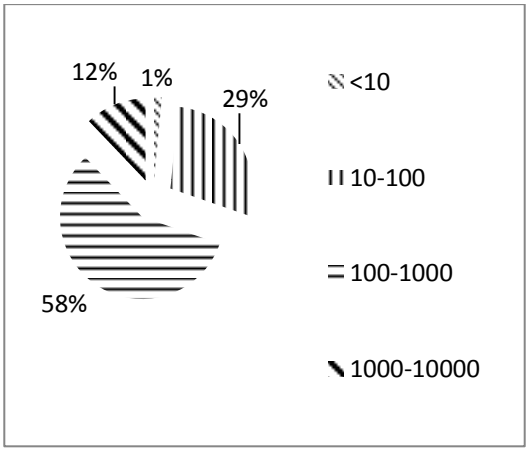


Figure 8: Number of Followers

The fourth question was “How often do you use Twitter?” which was measured on a 4-point Likert-type Scale. The scales were “Many Times a day”, ”About once a day”, ”A few times per week “and “Less than once per month”.

As shown in Figure 9, 41 users among 66 participants (62%) use Twitter more than once a day. In addition, 86% of our participants have login at least once a day.

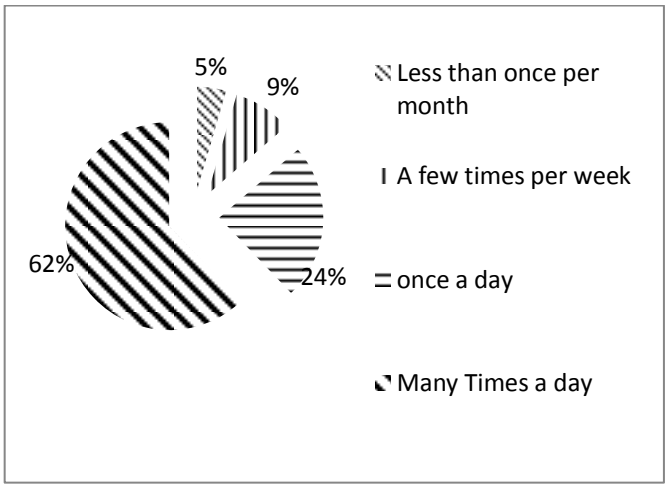


Figure 9: How Often Do You Use Twitter?

Figure 9 shows the distributions of the reason why they use Twitter. 33% of the respondents tend to use Twitter to get news and 19% of them use Twitter for having fun. As seen from Figure 10, the least important reason is following celebrities (4%).

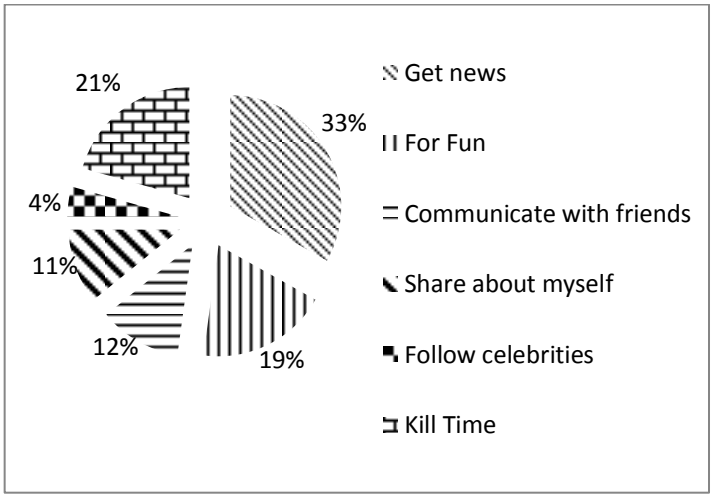


Figure 10: Why Do You Use Twitter?

4.2.2 Association between Facebook's "like" and Twitter's "favorite"

In order to describe one of the primary uses of the favoriting button, we try to find the association between Twitter's favorite button and the Facebook's "Like" button. As seen from in Figure 11, among 66 responses, 37 of them (56%) agree that favoriting in Twitter is similar with liking behaviour in Facebook.

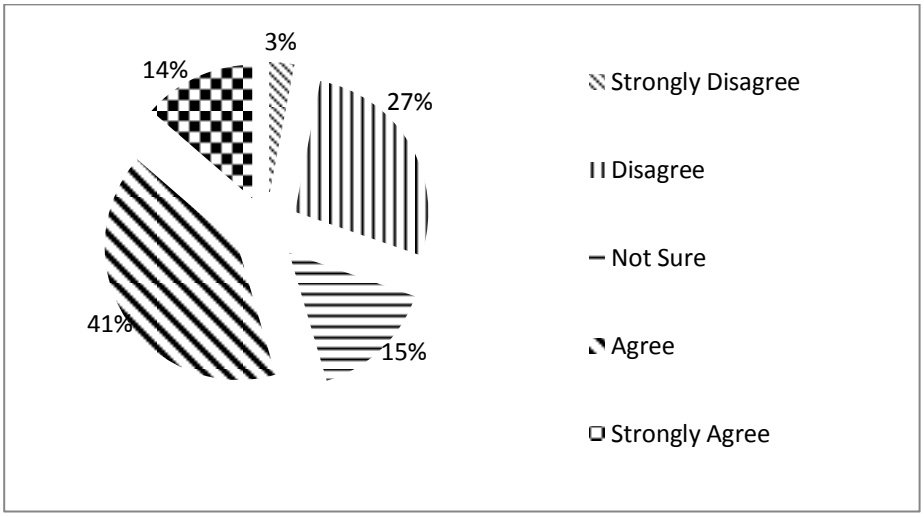


Figure 11: Similarity Between Facebook's "Like" and Twitter's "Fav"

4.2.3 Why people use favorite in Twitter?

One of the main aims of our study is finding the effects of using favorite in Twitter. According to our survey, pressing the favorite button is motivated by a range of different reasons as represented in Figure 12. In this survey item, we asked 10 possibilities that they might consider when they favorited a tweet. Possibilities were measured on a 5-point Likert-type Scale. The responses range from "Strongly Disagree" to "Strongly Agree". The majority of the participants (95%) tend to use favorite when they like it. This result is in compliance with the previous question regarding similarity between "like" button in Facebook and "favorite" in Twitter. 90% of the participants' favorite a tweet when they find it interesting and 81% of them favorite a tweet when the tweet is funny. On the other hand only 15% of participants use favorite for flirting someone and only 01% of the participants use favorite when they hate the tweet.

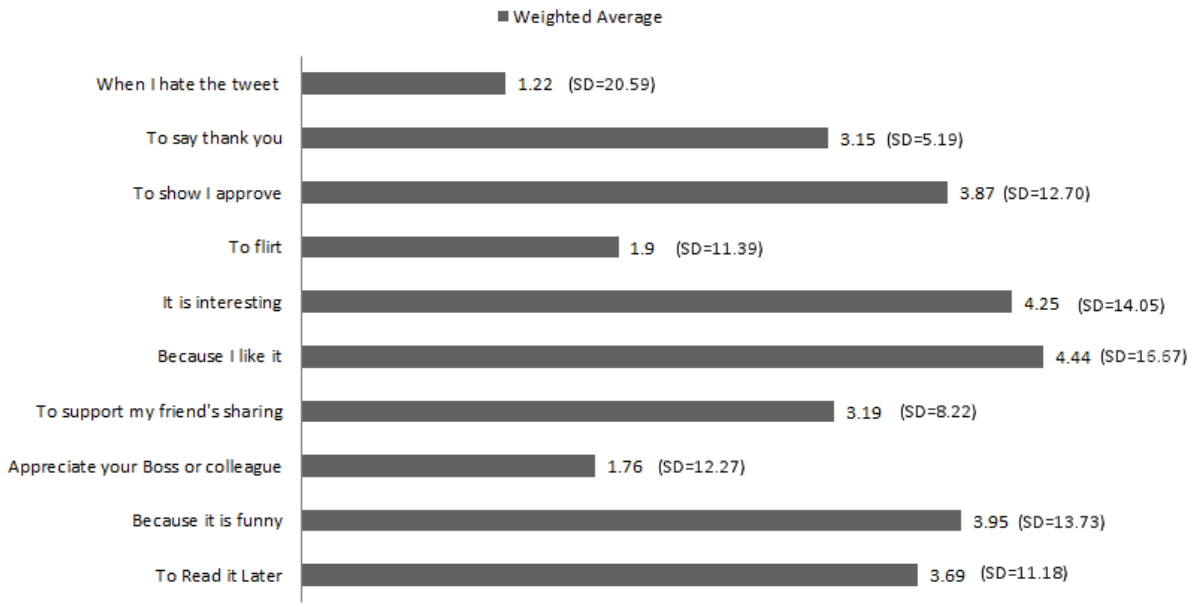


Figure 12: Why Do You Use Favorite?

4.2.3 Why people use retweet in Twitter?

In order to understand the reasons why people use retweets in Twitter, we asked 10 possibilities that our responders might consider while retweeting a tweet. Possibilities were measured on a 5-point Likert-type Scale. The responses ranged from "Strongly Disagree" to "Strongly Agree". As seen in Figure 13, 96 % of the responders agree that they retweet a tweet when they broadcast news or tweets. According to results, similar to favorite, people tend to use retweet when they like the tweet (77%). On the other hand, the majority of our responders do not tend to retweet a tweet to gain followers (89%) or to advertise (85%). 32% of our participants agree that they might retweet a tweet even though they hate the content. When we compare retweet and favorite behaviour in hatred retweets, the percentage in retweet (32%) is quite higher than favorite (1%).

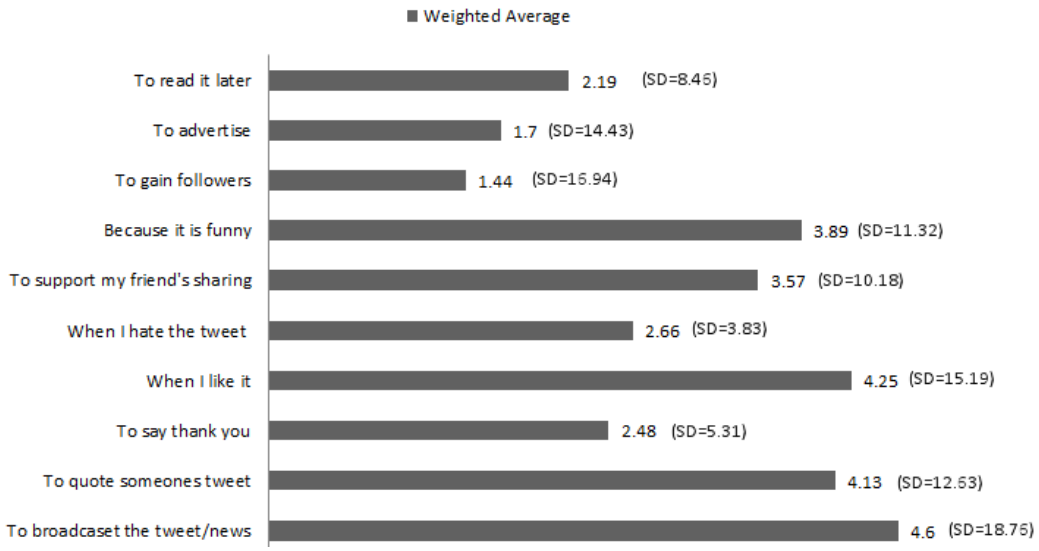


Figure 13: Why Do You Retweet a Tweet?

CHAPTER 5

METHODOLOGY

In this chapter we inspect Twitter from a recommender system point of view. The aim is developing a new recommendation approach to suggest new followees for Twitter users according to their interactions and content base similarities. In order to get more effective results, collaborative filtering and content-based similarity are used together. Firstly an overview of the system design is presented. Secondly, topological approaches are described. Lastly, content-based approach is described in detail.

5.1 System Design

Twitter is a directed social-graph which has 240 million users. Everyday 500 million tweets are posted that contain valuable information [35]. In this study, Twitter is used as a powerful data source for our recommender system. Our aim is finding desirable users to follow from this data. The architecture of the system is shown in Figure 14. The proposed system has two main parts; network topology and content analysis. In the network topology, we utilize relationship characteristics. Our objective is to benefit from the effectiveness of the relationships. We examine three different kinds of relationships from target user's neighbourhood; follow, retweet and favorite.

Studies show that [17] topological closeness has a positive effect on finding new people to follow on Twitter. In topology part of our study, we recommend new followees to our target users based on close relationships between users. For each target user, a neighbourhood topology is constructed with maximum path length 2, based on the followee - follower, retweet and favorite relationships.

In content analysis, our objective is to find the effectiveness of content in followee recommendation. In this part of the study, for each user top 100 ranked suggestions are chosen from the results of topological analysis and they are further enhanced with content analysis.

Since our system is based on topological and content data, in order to avoid problems data sparsity problems, we examine only the active Twitter users who have at least 50 followees/followers and publish at least 10 tweets per a month.

In content based analysis, two approaches are examined; topical similarity and opinion similarity. In topical similarity, topics from users' tweets are extracted and users are ranked accordingly the similarity between their topics and target user's topics. In opinion similarity, users' opinions are obtained by including the sentiment approach in the topical analysis.

In order to make better recommendations, some of the proposed strategies are merged. Firstly, topological approaches are combined to utilize the efficiency of topological methodologies. Secondly, both topological and content based approaches are merged to find the effects of content analysis on finding new people to follow on Twitter.

Finally, all recommendation methodologies are merged to find the best approach for a followee recommender system for Twitter. For each strategy, top 10 ranked users are chosen in order to evaluate whether the recommendations are relevant or not.

The details of the examined approaches are explained in the following sections.

For the implementation of the proposed method, Java[4] programming language is used while implementing core application and JSP[5] is used for GUI. NetBeans[6] is used as compiler. In data crawling part, Twitter RESTful API is used with the help of twitter4j [49]. MySQL[7] is used for storing data and recommendations.

[4] http://www.oracle.com/technetwork/java/index.html
[5] http://www.oracle.com/technetwork/java/javaee/jsp/index.html
[6] https://netbeans.org/
[7] https://www.mysql.com/

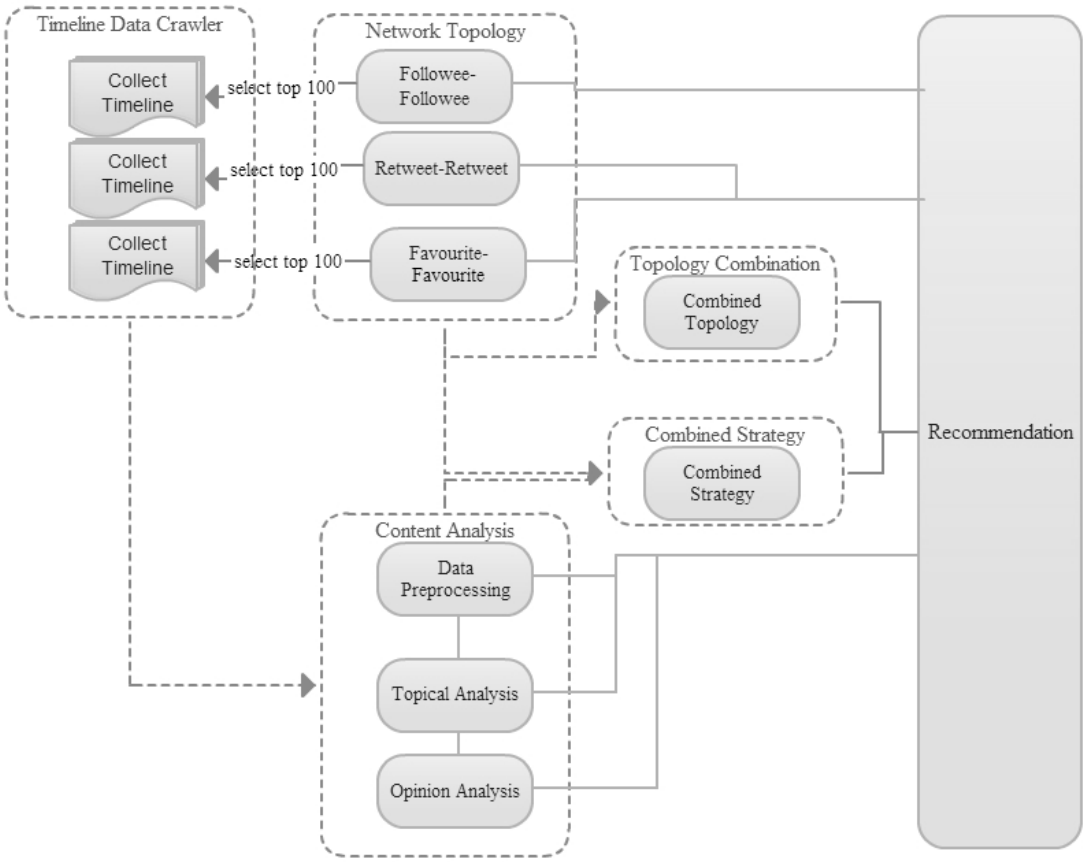


Figure 14: System Architecture

5.2 Topological Approach

Network topology is a substantial input for recommender systems. In Twitter, social network topology is formed by various directional relationships among users such as followee, follower, retweet and favorited.

When a user starts following a followee; it means that the user is willing to get updates from chosen followee's recent status. In addition, according to our survey, user's retweet a tweet when it is worth to share or they like it. Furthermore, favoriting shows that users appreciation on the tweet.

Trust is defined from different aspects in several studies [2, 50, 51, 52]. In this study, we use Golbeck's definition [50] which was applied for social network analysis in Twitter before [42], and states that "trust in a person is a commitment to an action based on belief that the future actions of that person will lead to a good outcome". On the basis of Golbeck's definition, "transitivity" is used in our topology based strategies. Figure 15 shows the transitivity relationship between User A, B and C, which means that if User A trusts User B and if User B trusts User C, then we could say that User A might trust User C as well.

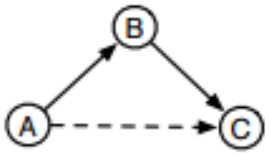


Figure 15: Transitivity Relationship

Although followee-follower relationships show the trust relationship between users, studies [31, 42] showed that users do not interact with all of their followees. Hence each relationship has diversity in quality, in order to differentiate the quality of the relationships between users; we also focus on the hidden relations between users such as retweets and favorites. In some studies [32, 42], these hidden relationships are also referred as web-of-trust or "circle-of-trust' [31].

In this section, we describe different types of relationships in more detail and we evaluate the effects of each relation type on the users' followee choices in Twitter.

5.2.1 Followees of Followees

In this part, user-followee relationships are used as a baseline for the recommendations. Recommendations are basically based on finding users from followees of User *u*'s followees. We assume that if User *u* follows someone, the followed user and our user have a trust relationship. On the basis of this *trust* relationship, if a user who has been followed by target user's followee can be assumed to be interesting for user u. Followees of followee network are represented in Figure 16. Users are represented as nodes and links between the nodes indicate the followee relationship between users.

For instance, in Figure 16, the link between User x and User y means that User x follows User y. We denote this relation as $x \in followee(y)$. This link at the same time it also means that User y is followed by User x. We denote this relation as $y \in follower(x)$.

In our system, each followee-user relationship is represented as a tuple that contain user id and user id of the followed user, as shown as follows:

Tuple = (user id, Followee id)

For instance, the followee relationship between User x and User y can be represented as a tuple as *(x, y)*.

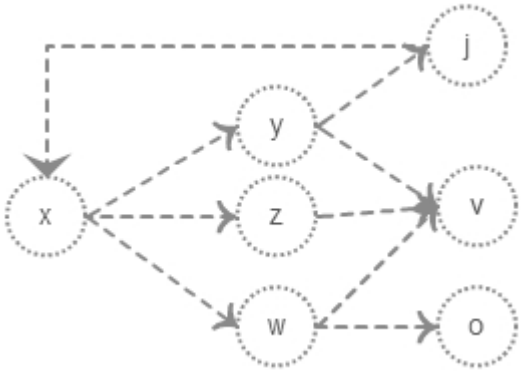


Figure 16: User-Followee Network Structure

More formally;

1) Let *u* be a user, we present *u*'s followee list as $E(u)$.

$$E\,(u) = \bigcup_{f \in followee(u)} (u, f) \tag{3}$$

2) Collecting followee data for each user f in $E(u)$. Let us call it $T(u)$.

$$T\,(u) = \bigcup_{f \in followee(u)} E(f) \tag{4}$$

3) Collecting follower list for User *u*. We present *u*'s followee list as $W(u)$.

$$W\,(u) = \bigcup_{w \in follower(u)} (w, u) \tag{5}$$

Firstly, in Equation 3, we get User *u*'s followee list $E(u)$. Secondly, in Equation 4 we get all followees for each user in $E(u)$. We build directed topology graph by getting followee list of each user in $E(u)$. Thirdly, in Equation 5, we get the followers of user *u*. As followers and followees are already known by our target user u, we eliminate the user's followers in list $T(u)$. Twitter does not allow us to get information about the protected users for that reason we eliminate the protected users from $E(u)$ and $T(u)$ in Equation 6. After elimination process, finally we are able to constitute the candidate list C_f.

4) Eliminating the followers

$$C_f(u) = T(u) \setminus (W(u) \cup E(u)) \tag{6}$$

This candidate list (C_f) is a set of tuples. Lastly, for each followee-id from the tuples, the number of occurrences is counted in the candidate list C_f and ranked according to

their occurrence count. Lastly, the top ranked 10 users from list $T(u)$ are selected in order to recommend to the user u.

For instance, the following steps show the example of recommendation process for user x from Figure 16 by using the formulas (Equation 3), (Equation 4), (Equation 5) and (Equation 6) respectively.

1) In Step 1, the followees of user *x* are collected.

   followee(x) ={y, z, w}

   E (x) = {(x, *y), (x, z), (x, w)*}

2) In Step 2, followee data is collected for each user from *E(x).*

   E (y) = {(y, *j), (y, v)*}

   E (z) = {(z, *v)*}

   E (w) = {(w, *v), (w, o)*}

   After collecting followee data for each user, T(x) is generated.

   T(x) = E (y) U E (z) U E (w)

   Therefore, T(x) = {(y, *j), (y, v),* (z, *v),* (w, *v), (w, o)*}

3) In Step 3, the follower data of user *x* is collected.

   W (x) = {(*j, x)*}

4) In Step 4, since User *x* is already aware of own followees and followers, we filter these users from the list. At the end of the filtering process, since the user *j* is a follower of User *x*, User *y* is removed from *E(y)*

   $C_f(x) = \{(y, v), (z, v), (w, v), (w, o)\}$

   For each followee relationship between the candidate users, tuples are created. $C_f(x)$ shows all the candidate tuples that are generated for User *x* from Figure 16, are listed in Table 3.

   As shown in Table 3, candidate User '*v*' is observed three times, candidate User '*o*' appears only once. From this list we can assume that, User '*v*' could be a strong candidate for our target user.

Table 3: Candidate Tuples that generated for User *x*

Candidate Tuples that generated from y node	*(y, v)*
Candidate Tuples that generated from z node	*(z, v)*
Candidate Tuples that generated from w node	*(w, v)* *(w, o)*

5.2.2 Favorites of Favorites

Favorites are one of the interaction ways in Twitter that show the trust relationship between users [31, 41, 42]. As stated in [41], Twitter users use favorites to collect interesting tweets as bookmarked tweets to be read later. In our dataset users tend to use favorites as much as retweets. The usage of favorites has been increased in the recent years. Our survey shows that users have started to use favorites similar to "like" button in Facebook. According to our survey, users do not hesitate to use favorites to appreciate the tweets in their close network. On the contrary to the other studies in the literature [41, 42], in this section we concentrate on favorites separately. Similar to followees of followees strategy, simple-transitivity is used while finding the candidates, which means that if a user favorites a tweet, this shows the trust relationship between the user and the author of the tweet. By using user's favorites as a base of trust relationship, we recommend users who have been most favorited by the users that the target user favorited.

5) Let *u* be a user, we present *u*'s favorited user list as $M(u)$.

$$M(u) = \bigcup_{fav \in favourite(u)} (u, fav) \tag{8}$$

6) Collecting favorited tweets for each user favorited in *M*. Let us call this set as $S(u)$.

$$\mathrm{S}(u) = \bigcup_{fav \in \mathrm{M}(u)} M(fav) \tag{9}$$

7) Eliminating the followers and followees

$$C_{fav}(u) = S(u) / (\mathrm{W}(u) \cup (u)) \tag{10}$$

Firstly in Equation 8, the user *u*'s favorited tweets are collected and the authors of these tweets are selected. We create a favorited user list $M(u)$ that represents the user-favorite relationships. In Equation 9, all favorited tweets are collected from each

user in $M(u)$. After eliminating followers, followees and protected users Equation 10, we form up a directed favorite user graph that has a maximum path length of two and centred by target user. We count every occurrence of each user and select most popular 10 favorited users from the candidate list C_{fav} to recommend to user u.

5.2.3 Retweets of Retweets

Retweeting is one of the communication ways that shows the trust relationship between people [18]. According to our survey results, despite some minor exceptions (hatred tweets), when a user retweets a tweet, it shows that the information in that tweet is interesting and worth to share.

In this section, we use retweet as a trust relationship between users and this trust relationship is used as a baseline of simple transitive recommendations. We recommend a user who has been retweeted by someone that our user retweeted before.

8) Let u be a user, we present u's retweet user list. Let us call this list $R(u)$.

$$R\,(u) = \bigcup_{r \in retweet(u)} (u, r) \tag{11}$$

9) Collecting retweeted tweets for each user r in $R(u)$. Let us call it $RT(u)$

$$RT\,(u) = \bigcup_{r \in R(u)} R(r) \tag{12}$$

10) Eliminating the followers and followees

$$C_{rt}(u) = RT(u) \,/\, (W(u) \cup E(u)) \tag{13}$$

Equation 11 shows the collection of User u's retweets from user u's timeline and a user list is created from the creators of these retweets. In Equation 12, we repeat Step 8 for each user in retweet list $R(u)$. Since Twitter does not allow us to get information about protected users, we also eliminate the protected users from $RT(u)$ list. We assume that followers and followees are already known by our target user u. For this reason, in Equation 13 we eliminate user's followers and followees in the list $RT(u)$ in Step 10. After elimination process, we finally constitute the candidate list C_{rt}.At the end of Step 10, we build a directed retweet graph that has a maximum path length of 2 and centred by target User u. Finally, we count the number of occurrences for each user in list. We choose up to 10 users who are the most retweeted users among all users C_{rt}.

5.3 Content Analysis

Previous studies [4] showed that shared content has significant effects on follower growth in Twitter. Being topically focused, sharing retweetable, informative contents, using hashtags and sentiments of tweets have considerable impacts on

gaining followers. Another study [43] shows that people tend to follow users who have similar interests or similar topics.

As stated before, in our graph-based topological approach, the effects of relationship types are tested. In order to improve our recommender system, we include the content approach to our methodology.

Our content based approach has two phases; topic similarity and opinion similarity. For content analysis, top 100 users from our graph-based approach candidate lists are chosen from each strategy (Followees of Followees, Retweets of Retweets, and Favorites of Favorites). We collected their timelines for 75 days. In the following discussions, we use this collected timeline data for analysing contents.

5.3.1 Data Collection

In this study, as mentioned in previous phases, Twitter Rest API is used in order to collect data from Twitter with help of twitter4j java library. MySQL DB is used to store data. As stated before, at the end of our network structure based analysis, 100 possible users are obtained from first three strategies for each target user. We collect all target users and first three strategies top listed 100 users' time line date and users' profile data.

5.3.2 Data Pre-processing

Topic analysis in Twitter is quite challenging because of the specific characteristic of Twitter data. Twitter is a free space to share information in every form. Users mostly use informal language and do not have to follow grammatical rules. Additionally, not all tweets have meaningful content. Thus elimination of noisy data is essential for extracting more meaningful topics from unstructured data. In this section, we describe the data pre-processing steps that are applied in our system.

5.3.2.1 Extracting Features from Tweets

Twitter enables sharing different kinds of data for several purposes. For example a status can contain a link which indicates a picture, location or a reference to an external website link and in addition mentions (@username) are also used to address a user in a status. Since these data do not refer any topical content, tweets are cleaned up in this process. Figure 17 shows the data pre-processing design steps that we apply in order to filter data.

Figure 18 shows an original tweet written in Turkish that includes mention, link and hashtag. Before starting topic analysis, for each tweet in our database, we filter the data to eliminate these features. Table 4 shows the filtering steps. In Step 1, links are removed from tweets. In Step 2, username starting with '@' character is removed from tweets. We remove all punctuations including emoticons and # in Step 3. All punctuations are removed in Step 4. Moreover we remove all tweets that include fewer than three words since they do not have enough topical keywords to extract topics from it.

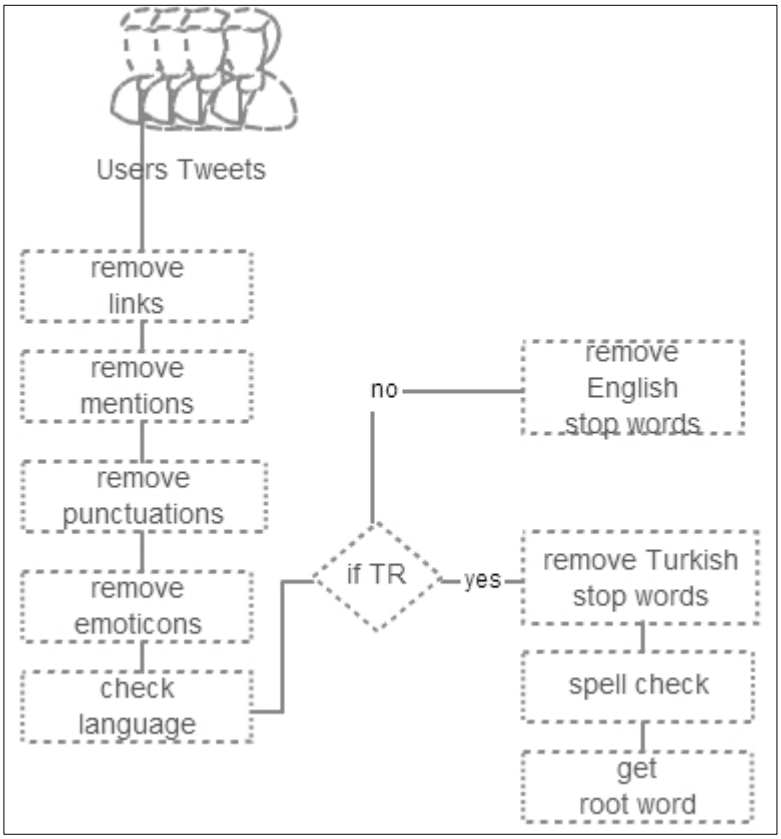


Figure 17: Data Pre-processing

Figure 18: Sample Original Tweet

Table 4: Data Pre-processing

1	#yfyi 2014 final torenine katilan tum gruplara tesekkur ediyor, odul kazanan tum gruplari tebrik ediyoruz. @yfyi
2	#yfyi 2014 final torenine katilan tum gruplara tesekkur ediyor, odul kazanan tum gruplari tebrik ediyoruz.
3	yfyi 2014 final torenine katilan tum gruplara tesekkur ediyor, odul kazanan tum gruplari tebrik ediyoruz.
4	yfyi 2014 final torenine katilan gruplara tesekkur ediyor odul kazanan gruplari tebrik ediyoruz
5	yfyi 2014 final törenine katılan tüm gruplara teşekkür ediyor ödül kazanan tüm grupları tebrik ediyoruz
6	yfyi 2014 final törenine katılan gruplara teşekkür ediyor ödül kazanan grupları tebrik ediyoruz
7	yfyi 2014 final tören katıl grup teşekkür et ödül kazan grup tebrik et

5.3.2.2 Language Detection

Our proposed system works on both English and Turkish texts. The content analysis process works differently for English and Turkish. Zemberek [12] is a comprehensive Turkish NLP framework, which has a database containing over one million Turkish words. Zemberek [12] has the capability to detect the sentences in Turkish language. It recognizes Turkish and non-Turkish words in a sentence and returns results according to the proportion of the Turkish words. Figure 19 shows the pseudo code of the language detection process. “languageTest()" is a method that is provided by Zemberek [12]. “languageTest()” method returns a string value that indicates the proportion of the Turkish words in the sentence. These return values are NONE, LESS LIKELY, LIKELY and MOST LIKELY.

In our study, we use LESS LIKELY, LIKELY and MOST LIKELY sentences as a Turkish phrase. After Step 4 in Table 4, we check the language of the sentence and the result of Zemberek indicates that the sentence is most likely written in Turkish.

```
TurkishTextDetection(String sentence) {

        result = languageTest(sentence);

      if result is NONE {
         return NOT-TURKISH;
      }
      if result is LESS LIKELY {
         return LESS LIKELY TURKISH
      }
      if (result == LIKELY) {
         return  LIKELY TURKISH
      }
      if (result == MOST LIKELY) {
         return MOST LIKELY TURKISH
      }
   }
```

Figure 19: Pseudo Code for Language Detection

5.3.2.3 Spell Checking

Since Twitter does not require any grammar rule, spellchecking helps to improve content analysis in Twitter. Turkish has special letters like "ç, ı, ğ, ö, ş, ü". A word can be written in a wrong way without using these characters in Twitter. Zemberek [12] is used for spellchecking for Turkish text. It provides a converter from ASCII to Turkish for spellchecking in Turkish. For example, as shown Step 5 in Table 3, after spell checking, "tesekkur" is corrected to "teşekkür".

5.3.2.3 Stop Word Removal

Stop words are short function words that are generally non-informative, such as conjunctions, numbers, pronouns etc. Stop word removal helps us to get more meaningful and clean topics. Since our system works both for English and Turkish, each language has its own stop word list. After language detection, stop word removal is accomplished according to the language. As shown Step 6 in Table 4, stop word "tüm" is removed from Step 5. Our stop word list includes some Twitter specific words such as RT, fav, TT, mention. Stop word lists for both languages are listed in Appendix B-1 and B-2.

5.3.2.4 Stemming

As we mentioned above, since Turkish is an agglutinative language, a word could be formed by derivations of suffixes. LDA is based on word frequency and combinations of the words in a tweet. If a word has different versions due to the suffixes, the chance of finding common phrases from the word corpus would be harder. The morphological ambiguity of the words makes the topic analysis less effective. For finding more possible combinations of words in Turkish, the roots of the words are extracted. Zemberek [12] is used in order to analyse the language of the tweet and to find root of the word. Figure 20 shows the suggestions that are provided by Zemberek for the word "katılan". For some words, like hashtags,

Zemberek cannot find the root. Hashtags could be combination of more than one word, thus we leave the word as it is. In Step 7, final state of the sample sentence is seen after the stemming process.

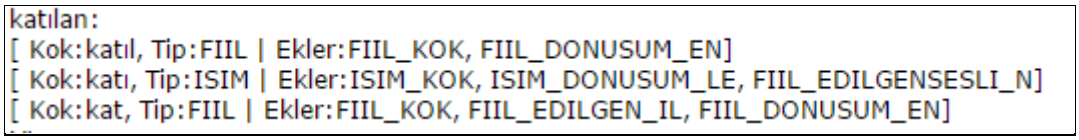


Figure 20: Stemming Suggestions from Zemberek

5.3.2 Finding Topics

In our proposed system, Twitter-LDA is used for finding topics from the pre-processed data. As stated before, the main assumption in Twitter-LDA is that, one tweet is usually about only one topic. When a user writes a tweet, first of all, a topic is chosen based on the user's interests. Then words are selected related to the specific topic. It is assumed that not all the words in a tweet have topical meaning. There are some words that are used commonly, assumed as background words.

While applying Twitter-LDA to our system, we refer Zhao et al. [13] to determine Dirichlet distribution parameters' values. For T number of topics, we set $\alpha=T/50$, $\beta=0.01$ and $\gamma=20$. According to the collected data, we try a range of values and we found that a number between 70 and 80 was a good choice for our data set. The topic count is assumed as 75 and the count of background words which are high frequency common words, is assumed as 50. For our topic modelling system, we run 200 iterations of Gibbs sampling. Topics are assigned based on word-topic distribution in the tweet. Equation 14 shows the topic-word assignment process, $\phi_{t,\omega}$ showing the probability of the word $\boldsymbol{\omega}$ for topic model ϕ_t and $n^{\omega}_{t,y=1}$ showing for a topic t, the count of w, is sampled as a topical word. According to the $\emptyset_{t,\boldsymbol{\omega}}$, words is assigned to the topics. Equation 15 indicates the user-topic distribution. n_u^t indicates that the number of topic t is sampled by the user u and $\sum_{t=1}^{T} n_u^t$ shows the number of tweets that has topical content that is published by the user u. Second row shows the number of each topic count that is seen in the user's timeline. The total count of topics for a user is calculated by the equation $\sum_{t=1}^{T} n_u^t$.

At the end of the topic modelling process, our word corpuses had 5515926 words that were collected from 4.352 users' timeline for 75 topics. 20 topical words are listed for each topic. Examples from topical words for some topics are listed in Table 5 and some of the background words are listed in Table 6. Moreover, Figure 13 shows some of the analysed tweets from a sample user's timeline. In Table 7, the numbers that are placed after words "/T_n", indicates the topic number that specific the word belongs to. "/false" indicates that word belongs to the background word list. As you can see from in Table 7, each word is assigned to a topic list or the background word list.

$$\emptyset_{t,\omega} = \frac{n_{t,y=1}^{\omega} + \beta}{\sum_{\omega=1}^{V} n_{t,y=1}^{\omega} + V\beta'} \quad (14)$$

$$\theta_{u,t} = \frac{n_u^t + \alpha}{\sum_{t=1}^{T} n_u^t + Ta'} \quad (15)$$

Table 5: Sample Topic Word List

Topic	**Keywords**
Topic 1	good, lol, people, shit, read, man, day, yea, back, system, today, work, great, makes, public, time, nice, cc, save, people
Topic 2	eder, teşekkür, ol, allah, iç, hadi, ay, günaydın, hayır, arv, iyi, abi, çay, güzel, sev, gecele, rahat, zaman, sağol, inşallah
Topic 3	parti, chp, söyle, ak, genel, erdoğan, kurultay, eleştiri, çalış, yeni, pm, sosyal, kk, seçim, koyun, yılmaz, dönem, karşı, tıp, anahtar

Table 6: Sample Background Word List

ol	gör	sev
yap	yaz	çocuk
gel	güzel	önce
al	abi	kal
iyi	gün	geç
adam	yeni böyle	yine
git	zaman son	çık
ver	iste	ara
insan bak		

Table 7: Sample Analysed Tweets

Topic	**Analysed Tweet**
z=26	hapis/26 zor/false insan/false akli/26 yitir/false ilker/26 basbug/26 nun/false darbe/26 magduruyum/26 hapishane/26 sistem/false adina/26 yurek/26 burk/26
z=33	haziran/33 hollanda/false ispanya/false maci/33 sirasinda/false dunya/33 evin/false gir/false ceyrek/33 taktigim/33 cift/33 yapiyor/33 acaba/false mutlu/false tum/33 dert/false
z=37	erdoğan/37 uzun/false adam/false cumhurbaşkanı/37 yap/false küçük/false elle/false dua/false

5.3.2.1 Subtopic Elimination

People use Twitter for different purposes such as sharing personal information, personal conversations or latest news or events. According to Dann et al. [48], people use Twitter mostly to share personal information about themselves such as "Good morning!", "Lovely day!" or "@Mary nice pic!". Although these tweets are shared publicly, generally third parties are not interested in these non-topical tweets. Since the aim of our study is to find topical similarities between users, we eliminate non-topical content to gain more accurate topical similarities between users. As given in Table 5, topics 1 and 2 have non-topical keywords; we get topic 1 and topic 2 out of our topic list. After eliminating non-topical content, we have got 71 topics listed.

5.3.2.1Constructing Topic Vectors

While constructing topic vectors for each user, we follow the same methodology which is used by Zhao et al. in [13]. In Table 8, the first column shows the Twitter user id which is provided by Twitter and the second column shows the distribution of each topic for one user. For each user *u*, a topical vector is constructed with the distribution values of topics. User-topic distribution is shown in Equation 16, z_{t1}^{u1} corresponds to the distribution of topic *t1* on user's topical tweets.

$$\boldsymbol{C_K(u1) = \{z_{t1}^{u1}, z_{t2}^{u1}, z_{t3}^{u1}, z_{t4}^{u1} \ldots\}} \tag{16}$$

Table 8: Topic Distributions on Users

User Id	**Topic Distributions On Users**
197009434	0.022 0.019 0.004 0.026 0.048 0.026 0.022 0.004 0.004 0.033 0.055 0.015 0.015 0.004 0.011 0.019 0.022 0.026 0.019 0.022 0.026 0.030 0.052 0.022 0.011 0.055 0.033 0.015 0.062 0.030 0.011 0.015 0.004 0.033 0.038 0.015 0.026 0.022 0.041 0.022

5.3.3 Sentiment Analysis

In addition to the topical analysis, we include sentiment analysis to our study in order to find people's opinion similarities and analyse the effect of opinion similarities on followee gaining.

In our study, we use SentiStrength [16, 47] as a sentiment analysis tool that supports both English and Turkish.

Figure 21 shows the system design of sentiment joint topic analysis model. In our proposed system, while finding sentiments of tweets, firstly links, mentions and punctuations are removed from the tweets after which the sentiments of the tweets is calculated according to language. Since emoticons are important for people to express their feelings, emoticons are removed after the sentiment analysis. After

calculating sentiments, punctuations and stop words were removed also. Figure 21, shows the sentiments values from sample tweets

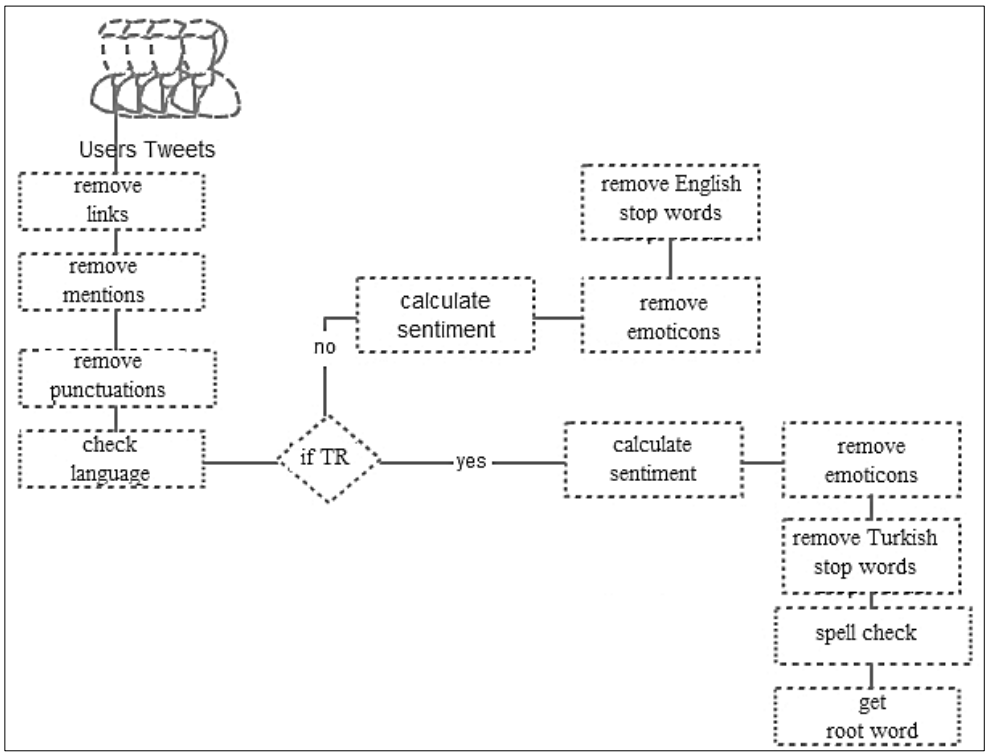


Figure 21: Topic- Sentiment Analysis System Design

5.3.3.1Constructing Sentiment-Topic Vectors

In order to understand a user's opinion about a topic, the average sentiment value is calculated for each topic of each user. Equation 17, shows the calculation of average sentiment value, in which $s_{u,n}^{t}$ denotes the sentiment value of tweet t which is sampled by user u for n times.

As a result, each user has sentiment values for each topic distribution. Table 9 shows the user sentiment-topical distribution over topics.

Table 9: Sentiment-Topical Distribution over Topics

User Id	Topic Distributions On Users
197009434	(s:-1.5, t:0.030), (s:-2.0, t:0.042), (s:-6.0, t:0.010), (s:-1.7, t:0.017), (s:5.0, t:0.017), (s:1.0, t:0.008), (s:-0.4, t:0.009), (s:6.0, t:0.009), (s:3.0, t:0.013), (s:-1.0, t:0.021), (s:-1.0, t:0.019), (s:-1.5, t:0.020), (s:4.0, t:0.004), (s:-4.0, t:0.017), (s:2.3, t:0.024), (s:1.6, t:0.024), (s:0.0, t:0.009), (s:1.0, t:0.015), (s:5.2, t:0.0163), (s:2.5, t:0.0186), (s:-4.0, t:0.020), (s:-4,7, t:0.012), (s:2.0 t:0.026), (s:0.8, t:0.010), (s:-2.1, t:0.031), (s:-2.0, t:0.010), (s:1.0, t:0.020), (s:-1.0, t:0.012), (s:0.0, t:0.016), (s:-1.0, t:0.075), (s:-1.3, t:0.008), (s:-2.3, t:0.023), (s:3.0, t:0.008), (s:4.0, t:0.010), (s:-3.6, t:0.014), (s:5.6, t:0.016), (s:2.0, t:0.044), (s:1.6, t:0.013), (s:2.8, t:0.015), (s:-1.6, t:0.016), (s:-3.1, t:0.013), (s:0.0, t:0.023), (s:-1.0, t:0.041), (s:4.2, t:0.018), (s:1.0, t:0.028), (s:-1.0, t:0.019), (s:2.0, t:0.027), (s:0.0, t:0.016), (s:3.3, t:0.029), (s:1.0, t:0.016)

$$\boldsymbol{average\ sentiment\ value} = \frac{\sum_{n=1}^{N} s_{u,n}^{t}}{n_u^t} \tag{17}$$

For each user *u*, senti-topical vector is constructed by using the distribution values of topics. User- topic vector is shown in Equation 18, $z_{t1,s1}^{u1}$ representing the distribution of topic *t1* with average sentiment value on user's topical tweets.

$$\boldsymbol{C_O(u1) = \left\{z_{t1,s1}^{u1}, z_{t2,s2}^{u1}, z_{t3,s3}^{u1}, z_{t4,}^{u1} \ldots\right\}} \tag{18}$$

5.3.4 Similarity Calculation

Each user has an opinion-user or topic-user distribution vector. To calculate opinion similarity between users, we use Euclidean distance.

Euclidean distance measures the distance between two vectors. More formally, as seen in Equation 19, $d_{euclidean}^{(x,y)}$ indicates the distance between x and y, n-dimensional vectors which are represented as $x = \left\{z_{t1,s1}^{x}, z_{t2,s2}^{x}, z_{t3,s3}^{x}, z_{t4,s4}^{x} \ldots\right\}$ and

y =$\left\{z_{t1,s1}^{y}, z_{t2,s2}^{y}, z_{t3,s3}^{y}, z_{t4,s4}^{y} \ldots\right\}$ for opinion-user vector and $x = \{z_{t1}^{x}, z_{t2}^{x}, z_{t3}^{x}, z_{t4}^{x} \ldots\}$ and y =$\left\{z_{t1}^{y}, z_{t2}^{y}, z_{t3}^{y}, z_{t4}^{y} \ldots\right\}$ for topic-user vectors.

$$d_{euclidean}^{(x,y)} = \sqrt{\sum_i (x_i - y_i)^2} \quad (19)$$

5.4 Combined Strategies

In order to enhance the efficiency of our proposed methodologies, we combine the strategies. As mentioned before, top ranked 100 users are selected from all approaches according to the number of occurrences of the candidate user in particular list. Before combining values, we normalize all values in every candidate list.

In this section, firstly we give general information about the normalization method that we use in our system. Secondly, we describe our combined methodologies.

5.4.1 Normalization

Data is normalized in order to ensure that the relative magnitude is meaningful. Since our recommendations are user specific, normalization will not affect the suggestion list. In the proposed method, each strategy has independent numeric scalar value that we have normalized in the range of [1, 2]. Unity-based normalization (Equation 20) is used for normalizing values in each strategy. In Equation 20, *x* shows the normalized value, x_{min} represents the lowest value in parameter space and x_{max} represents the highest value in parameter space.

$$\boldsymbol{Normalized\ Value} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (20)$$

5.4.2 Combination of Topical Approaches

In topological approaches section, we work on different kinds of relationships in Twitter. We propose three strategies with using followee- follower, favorite and retweets relationships in Twitter. In order to make better recommendations, we combine these three strategies.

As mentioned above, we normalize all the values in the candidate lists. After normalizing the values, as shown in Equation 21, we add three values and calculate the average values of them. Lastly, we choose top 10 people who have the highest values to show our target users.

$$\boldsymbol{N_{C_{tc}(u)}} = \frac{N_{C_f(u)} + N_{C_{fav}(u)} + N_{C_{rt}(u)}}{3} \quad (21)$$

5.4.3 Combination of all Approaches

In this study, our aim is finding valuable users from Twitter data. In order to find the users, we try different aspects. Firstly we try to get benefit from the effectiveness of the relationships. After selecting topologically closer users, we collect their timelines and find content based similarities. We propose five different basic approaches for followee recommendation.

To make more efficient recommendations, we decide to combine the values that we gain from our previous approaches. Since we calculate the normalized values for combined topological approaches ($N_{C_{tc}(u)}$), in order to calculate combination of all strategies, we calculate the average of combined topological approach and opinion similarity values ($N_{C_o(u)}$). Equation 22 shows the calculation of this combined approach where $C_o(u)$ indicates the values that are gained from opinion analysis and $C_{tc}(u)$ indicates the combined topological approach's values. After normalizing the values, we calculate average of these two approaches. At the end of the calculation, top 10 candidates have been chosen to be shown to the user.

$$\boldsymbol{N_{C_{cm}(u)} = \frac{N_{C_o(u)} + N_{C_{tc}(u)}}{2}} \tag{22}$$

CHAPTER 6

EXPERIMENTS AND DATA ANALYSIS

In this chapter, we explain the experiments that are conducted to determine the effectiveness of the proposed approaches. Firstly, we give brief information regarding our proposed recommender system, Wootch. Secondly the details about recommender engine and our data set are explained. Thirdly, we present the results of the experiments and lastly we share our analysis and discussion.

6.1 Experiment Design

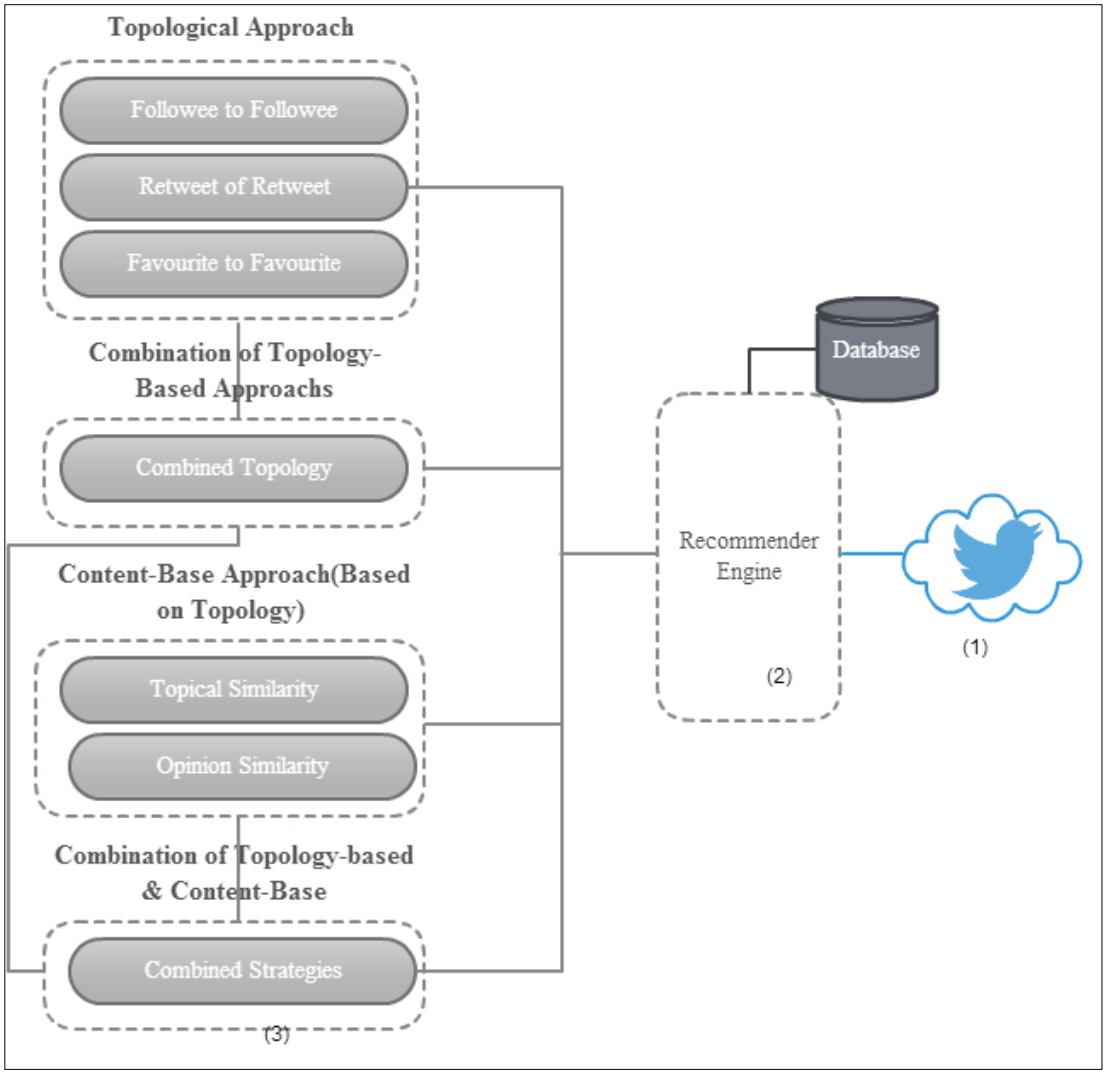


Figure 22: Wootch System Design

In order to generate recommendations and test our proposed strategies, we create a system called Wootch. As seen in Figure 22, in the first part, we collect data from Twitter. Secondly, in recommender engine, suggestions are generated. In the third part, recommended items are shown to our target users for evaluation.

In methodology section (Chapter 5), we explain our approaches which are listed in Table 10. For each proposed approach, a strategy number is assigned.

Table 10: Strategies

Strategy Number	**Strategy Type**	**Strategy Name**
Strategy 1	Topology-based	Followees of followees
Strategy 2	Topology-based	Retweets of Retweets
Strategy 3	Topology-based	Favorites of favorites
Strategy 4	Content-Base	Topic Similarity
Strategy 5	Content-Base	Opinion Similarity
Strategy 6	Combination of Topology-based	Topological Similarity
Strategy 7	Combination of Topology-based & Content-Base	Opinion Similarity + Topological Similarity

Figure 22 shows the system design. In the first part, topological strategies are tested in order to find the most effective topological approach. In the second part, content-based approach is applied on the basis of topological approach in order to enhance the effectiveness of recommendations. Lastly, previously mentioned strategies are combined: in Strategy 6, we combine Strategy 1, Strategy 2, and Strategy 3; where Strategy 7 is the combination of the topological and content based approaches. In Strategy 7, the results from different strategies (Strategy 1, Strategy 2, Strategy 3 and Strategy 5) are merged.

As mentioned in Chapter 5, in the first three strategies, we concentrate on topological aspects of the Twitter. In the first strategy, our recommendations are basically based on finding users from followees of our target user's followees. In the second strategy, we try to find the most favorited users on basis of target user's favorites. Lastly, in strategy three, most retweeted users are selected according to target user's retweets. At the end of the topology part, for each target user top ranked 100 users are collected from each strategy in order to use in content base analysis.

In order to find topically similar users, we start by collecting top 100 users' timelines which are gained from first three strategies. At the end of the timeline data crawling process, 250-300 users' timeline data are collected for each target user.

The proposed strategies are evaluated as follows: for each strategy, top 10 users are chosen from database and shown to the target user in order to rate recommended user for each strategy. A username/password combination is assigned for each user. When user logins to the system, the recommender engine chooses the top 10 personalized recommendations for each strategy to show to the user as seen in Figure 23.

The experiments are carried out for three weeks. 22 participants attend to these experiments. For each user the recommended item lists are presented as unordered list for each strategy. Firstly, we ask our participants to rank the four topological approaches at the same time and after collecting participant data for topology based algorithms; we ask them to rank content based approaches and combined methodology.

We also investigate the factors of rejecting items, as shown in Figure 24; we ask our users to clarify the reason/reasons why they do not want to follow the recommended user. The English translations of the presented reasons that are shown in Figure 24 are as follows:

1) I am not interested in this user's tweets.
2) This user posted so many tweets.
3) This user posted so many retweets.
4) I can reach this information from other sources
5) This user is not trustworthy.
6) This user is not sincere.
7) This user is repulsive. I do not like him/her.
8) Other

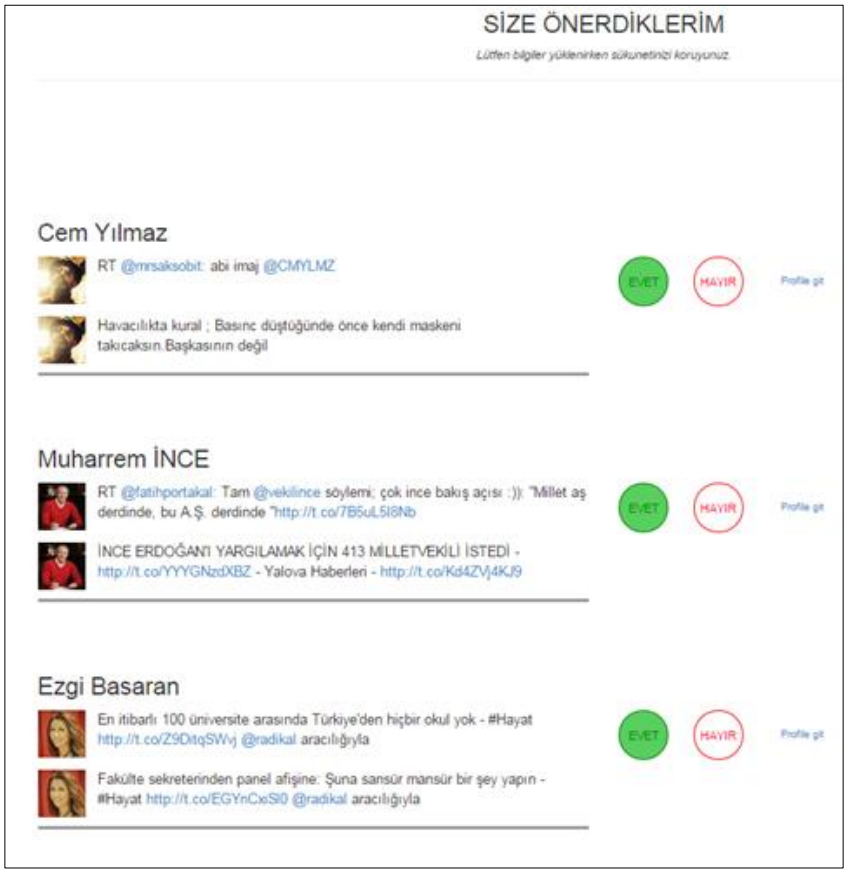


Figure 23: Recommender Engine

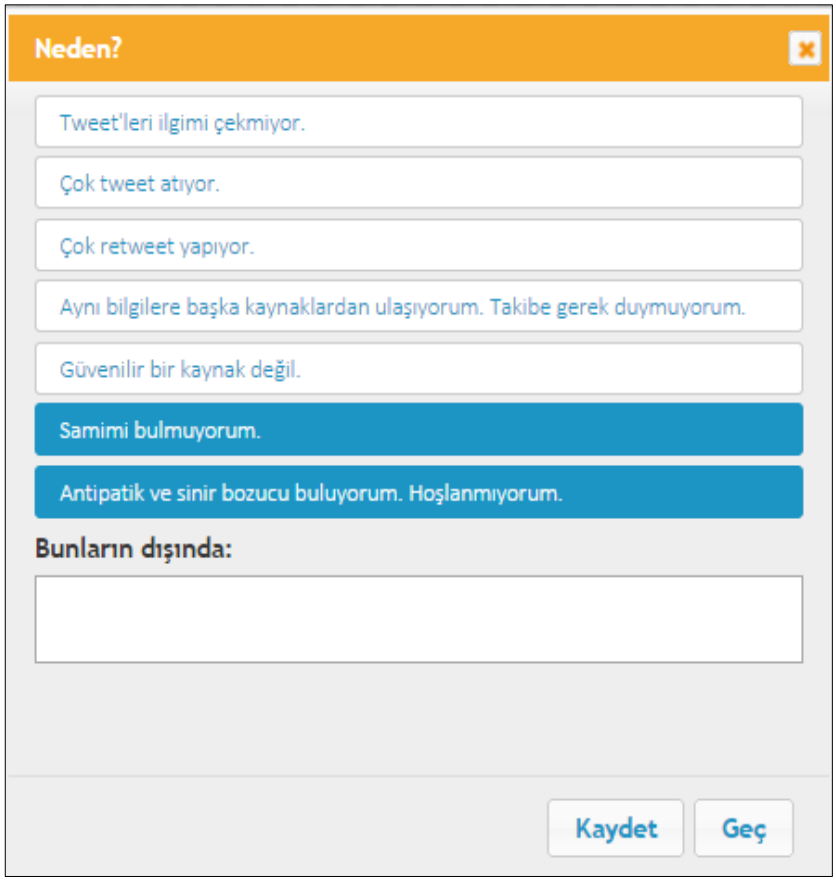


Figure 24: Why Don't You Want to Follow This User?

6.2 Data Gathering

As described in Section 6.1, the proposed strategies are mainly based on topological relatives between users. In order to find the topologically closer users, a user centred topological network graph is composed by using users' retweets, favorites and followees.

In data crawling process, we generate a simple topology for each user. Our data set includes the following data.

- Target users' followees
- Target users' followers
- Target users' followees of followees,
- Target users' favorites
- Favorited users' favorites
- Target users' retweets
- Retweeted user's retweets
- Timeline of top 100 users from each strategy.
  In order to evaluate the proposed strategies, we conduct an experiment with 22 users who use Twitter actively. Table 12 shows our attendees' followee-follower count and the number of tweets generated in their past two months. As seen from Table 12, participants have minimum 20 tweets / 10 favorites / 10 retweets and at least 20 followers / 50 followees.

Table 12: User Data

user id	followee #	follower#	2 month fav #	2 month RT #	2 month tweet #
user 1	280	82	133	61	79
user 2	188	33	190	13	116
user 3	251	135	53	27	93
user 4	191	180	155	186	198
user 5	155	187	50	69	123
user 6	573	252	5	26	40
user 7	298	192	154	27	16
user 8	71	90	97	41	127
user 9	327	201	56	89	69
user 10	209	140	21	10	20
user 11	305	168	38	104	80
user 12	175	229	126	290	200
user 13	482	240	55	31	35
user 14	588	435	11	16	40
user 15	809	1713	307	178	200
user 16	368	125	5	15	11
user 17	186	319	11	32	199
user 18	279	284	52	3	22
user 19	141	100	10	13	20
user 20	292	19	56	99	61
user 21	260	432	109	68	153
User 22	1059	757	1383	43	199
AVERAGE	340.3182	286.9545	82.09524	65.04545	94.59091
STD	235.0642	357.2559	286.824	71.48791	70.75892

The data has been collected from Twitter from January 2015 to March 2015. Over this period of two months, we have crawled 4.352 users with 2.806.429 followee-follower relationship and 453.607 favorited tweet, 213.642 retweeted tweet and 629,331 tweets.

6.3 Evaluation Metrics

Precision and recall are the basic measures that use in evaluating recommendation strategies. **Precision** shows the ratio of the number of relevant item (true positives) retrieved to the total number of irrelevant (false positives $\mathbf{F_p}$) and relevant items (true positives $\mathbf{T_p}$) retrieved (Equation 23). **Recall** measures the ratio of the number of relevant items (true positives) retrieved to the total number of relevant items (true positive $\mathbf{T_p}$ and false negative $\mathbf{F_n}$) in the set (Equation 24).

Average Precision (AP) shows the sum of precision value at each relevant item in recommendation list. In Equation 25, *k* indicates the rank in the sequence of recommended item list and *n* show the number of recommended item. *P(k)* shows the

precision value at position *k* and $\Delta r(k)$ shows the changed recall value between *k-1* and *k*.

Mean Average Precision (MAP) basically shows the average of AP which summarizes rankings from multiple users by averaging average precision (AP) values. Equation 26 shows the MAP formulation where *Q* indicates the number of recommended items. Although precision and recall give a general idea about the system performance, during precision and recall calculation, the order of recommended item is not used. For that reason, in order to determine the effectiveness of the ordered recommended item list, MAP evaluation (Equation 26) metric is commonly used in the recommender systems.

$$\mathbf{Precision} = \frac{\mathbf{T_p}}{\mathbf{T_p + F_p}} \tag{23}$$

$$\mathbf{Recall} = \frac{\mathbf{T_p}}{\mathbf{T_p + F_n}} \tag{24}$$

$$\mathbf{AP} = \sum_{k=1}^{n} \mathbf{P(k)\Delta r(k)} \tag{25}$$

$$\mathbf{MAP} = \frac{\sum_{q=1}^{Q} \mathbf{AP(q)}}{Q} \tag{26}$$

6.4 Experiments and Evaluation

Our proposed algorithms are evaluated in terms of their overall precision in followee recommendation. As stated before, mean average precision shows the number of relevant recommendation among all ordered recommended items. In order to find effectiveness of our algorithms in different positions, we rank our algorithms in three rankings: MAP@1, MAP@5, and MAP@10.

6.4.1 Topology-based Recommendation Experiments

6.4.1.1 Followees of Followees

In our first strategy, we evaluate our recommendations that are based on user-followee relationship. At the end of our experiment roughly 50% of our recommendations were relevant. Users agree to follow, 111 users out of 220 recommendations. Figure 25 shows the MAP values for positions 1, 5 and 10. As seen in Figure 25, the precision values are decreased in longer lists.

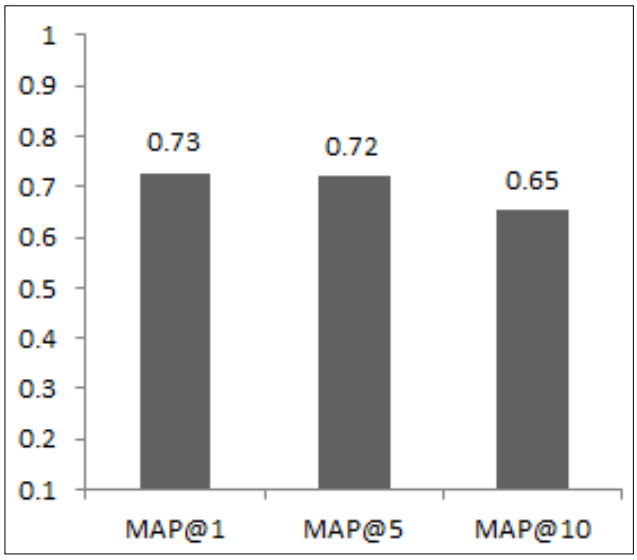


Figure 25: MAP Values for Followees of Followees Strategy

If users do not want to follow the recommended item, after the user clicks on "no" option, some possible reasons are listed as shown in Figure 26. The percentages of chosen reasons are given in Figure 23 for Strategy 1. The most significant reason is "I am not interested in this user's tweets", the second and third most popular reasons are "I can reach this information from other sources" and "This user is not sincere".

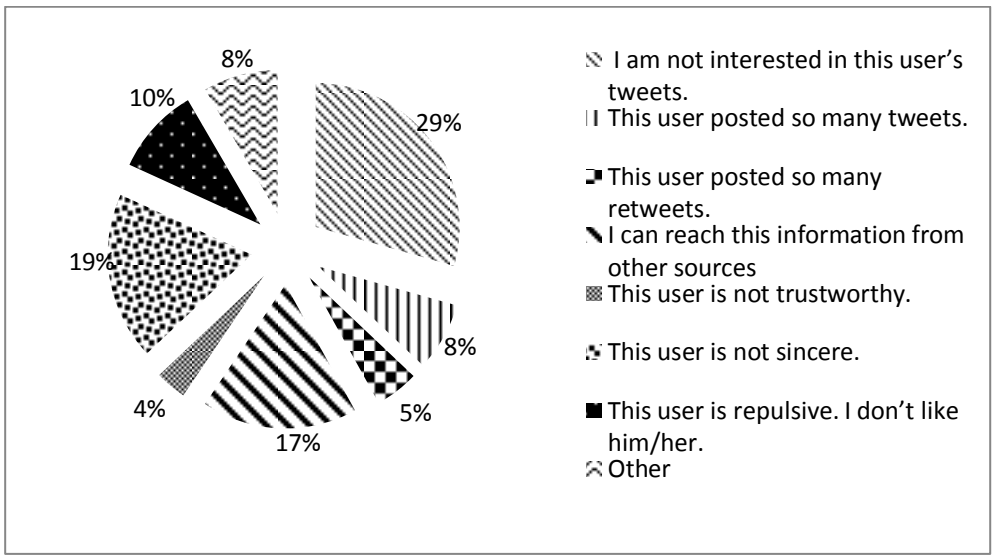


Figure 26: Disapproval Reasons for Followees of Followees Strategy

6.4.1.2 Favorites of Favorites

In this strategy, we rank our recommendations that are based on user's favorited tweets. At the end of the experiment 47% of our recommendations are relevant which means that the users agree to follow 104 users out of 220 recommended users.

Figure 27 shows the MAP values for position 1, 5 and 10. As the size of the ranking list increase, precision values increases.

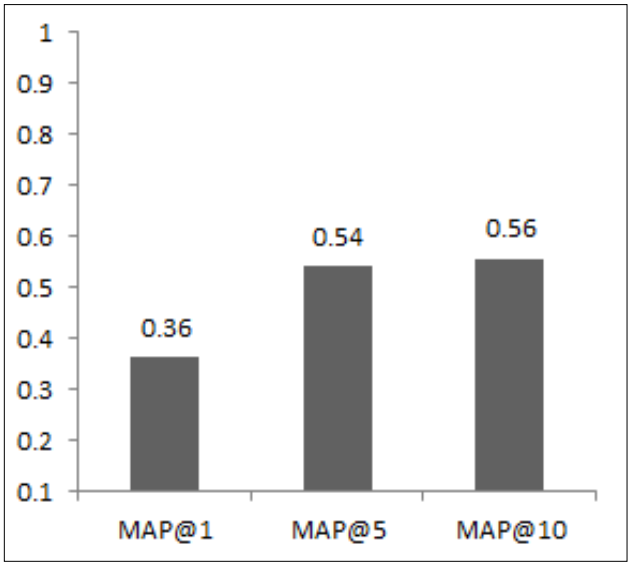


Figure 27: MAP Values for Favorites of Favorites Strategy

The percentages of the disapproval reasons are represented in Figure 28 for Strategy 2. The most significant reason is "I am not interested in this user's tweets" with 75% percentage, the second popular one is "I can reach this information from other sources".

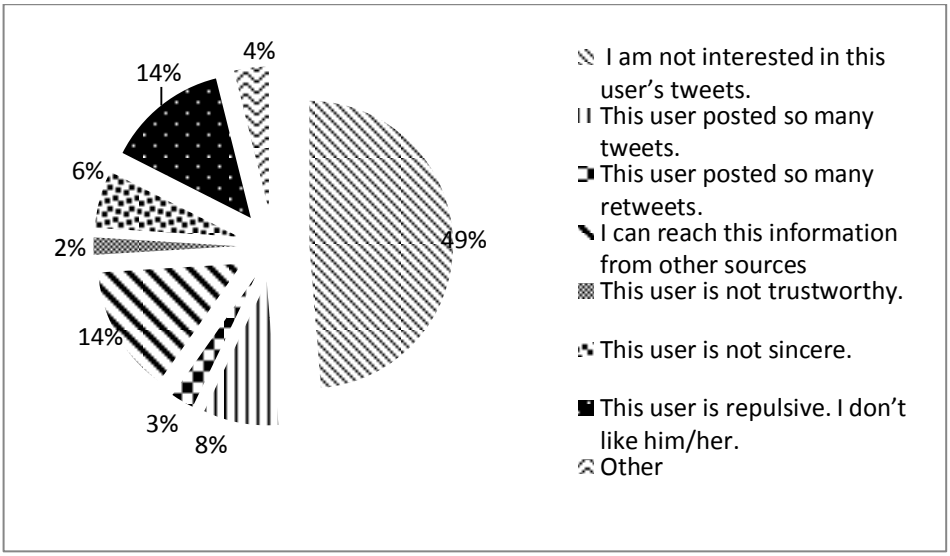


Figure 28: Disapproval Reasons for Favorites of Favorites Strategy

6.4.1.3 Retweets of Retweets

In this strategy, retweet based recommendations are evaluated. At the end of the experiment, the users accept 52% of our recommendations that means the users agree to follow 114 users out of 220 recommendations. Figure 29 shows the MAP values for the positions 1, 5 and 10. According to the results, MAP@5 gives the best performance.

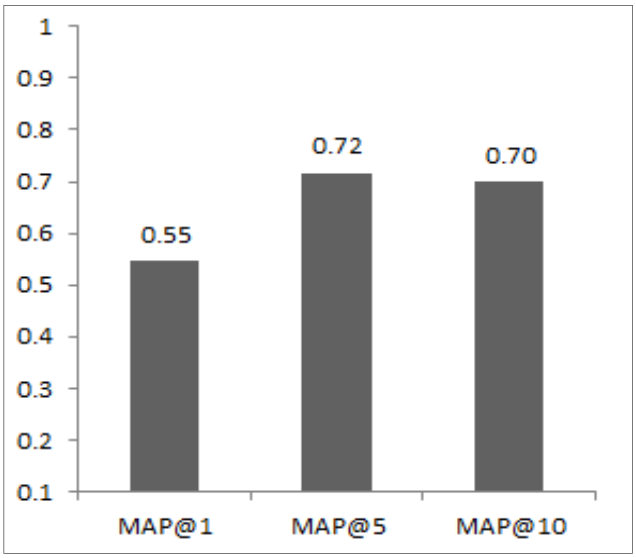


Figure 29: MAP Values for Retweets of Retweets Strategy

We can see the users' disapproval reasons in Figure 30 for Strategy 3. The most significant reason is "I can reach this information from other sources" with 39%, the second popular one is "I am not interested in this user's tweets" with 32%. Other than that, these two options are not significantly different from each other's.

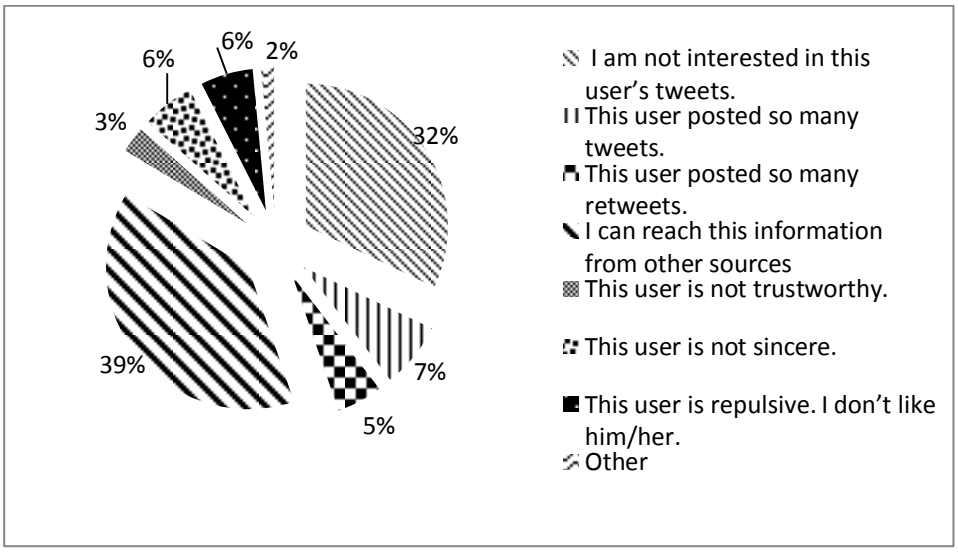


Figure 30: Disapproval Reasons for Retweets of Retweets Strategy

6.4.1.4 Experiment on Topology Combination Recommendation

In this experiment, we combine the proposed topological approaches that are followees of followees, favorites of favorites and retweets of retweets. At the end of the experiment, the users accept 67% of our recommendations, which means that the users agree to follow 147 users out of 220 recommended users. Figure 31 shows the results for MAP@1, MAP@5 and MAP@10. According to the results, MAP@5 best performance.

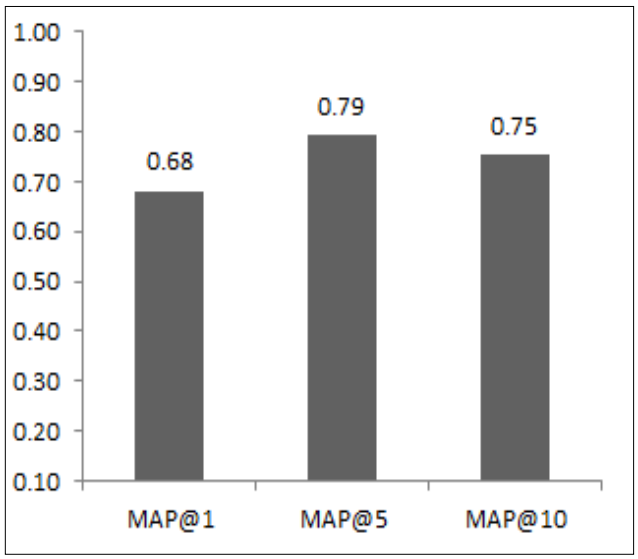


Figure 31: MAP Values for Combined Topology Strategy

As seen in Figure 32, the most popular rejection reason is "I am not interested in this user's tweets" with 44%. The second reason is with 21%, "I can reach this information from other sources".

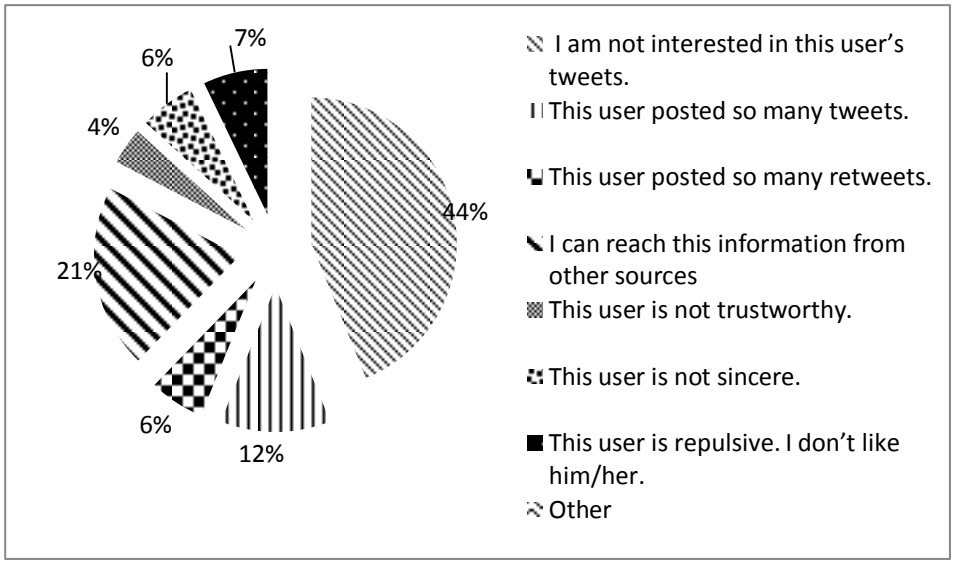


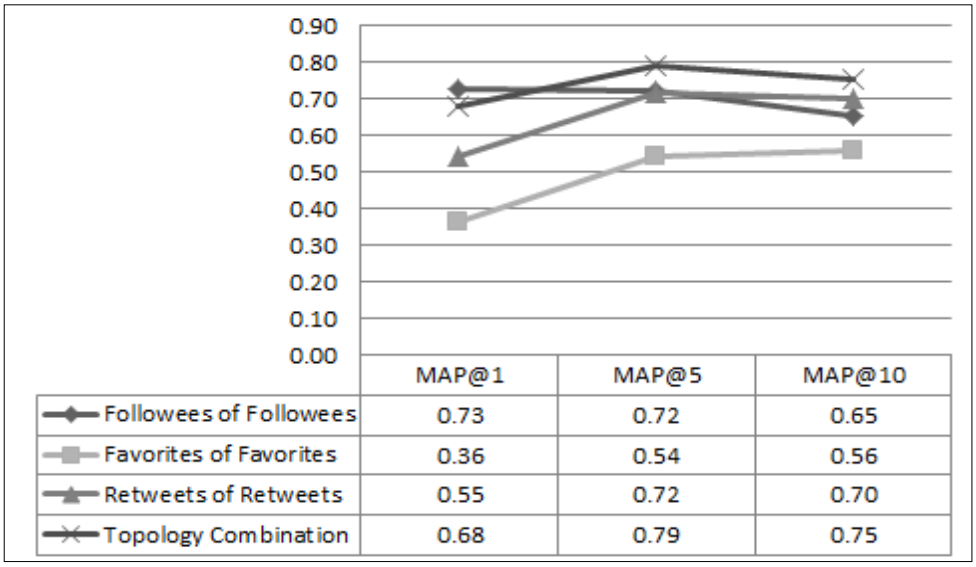
Figure 32: Disapproval Reasons for Combined Topology Strategy

6.4.1.5 Comparison of Topological Strategies

In topological analysis, we calculate the number of occurrences of a user in the list of recommendations that are gathered by using different Twitter characteristics. The comparisons of all topological strategies' results are presented in Figure 33. As seen in the figure, combined topology (Strategy 6) performs better than other strategies. In the second rank, Followees of Followees (Strategy 1) performs better than other topological strategies. Although at MAP@1, Strategy 1 performs better, at MAP@10, combined topology gives better results on the overall.

As seen in Figure 33, the best results are gained at MAP@5 and precision values decrease at MAP@10 in Strategy 2 and Strategy 3. We also find that Favorites of Favorites (Strategy 2) has the worst performance among other topological strategies.

Armento et al. [3] used a topology based algorithm for followee's recommender system for Twitter. They explored the graph of connections that is originated from the target user. They made a list of candidate users that they ranked using different weighting features such as popularity (#followers/#followees), the number occurrences of a given user in list of candidate users (w_c) and # of common friends. The number of occurrences strategy (w_c) is quite similar to our Strategy1. When we compare the results of ours with theirs experiments, their results are slightly different than ours. Their experiment results show that average precision values in MAP@1 is 0.9, on the other hand, that of our method are 0.72. The reason could be that our target users joined Twitter more than 2 years ago and they use Twitter actively on the other hand in their experiments Twitter users were newly joined Twitter.



	MAP@1	MAP@5	MAP@10
Followees of Followees	0.73	0.72	0.65
Favorites of Favorites	0.36	0.54	0.56
Retweets of Retweets	0.55	0.72	0.70
Topology Combination	0.68	0.79	0.75

Figure 33: Comparison of MAP Values for Topological Strategies

When we compare the Strategy 2 with the other strategies, although both strategies have high percentage on "I am not interested in this user's tweets" option (29%, 32%), the percentage of this item is relatively different in Strategy 2 (49%). We could deduce that people are disinterested in other user's favorites more retweets.

The percentage of "This user is repulsive. I don't like him/her" option has highest value in strategy 2(14%) and a lower value is seen in Strategy 3 with 6% percentage.

One of the most popular option in all strategies is "I can reach this information from other sources". Since retweeting is used as news propagation in Twitter, this option is significantly popular in Strategy 3 with 39% percentage. On the other hand Strategy 2 has lowest percentage with 14%.

6.4.2 Experiments on Content Analysis

Although we name this section content analysis, as explained on methodology part, content analysis is run on specific list of user's timeline data which gained from topological strategies.

6.4.2.1 Topical Similarity

In this strategy, we rank our recommendations that are based on topical similarities between the users. At the end of the experiment, the users agree to follow 109 users out of 220 recommended users. Figure 34 shows the mean average precision (MAP) values in @1, @5 and @10 where MAP@5gives the best results.

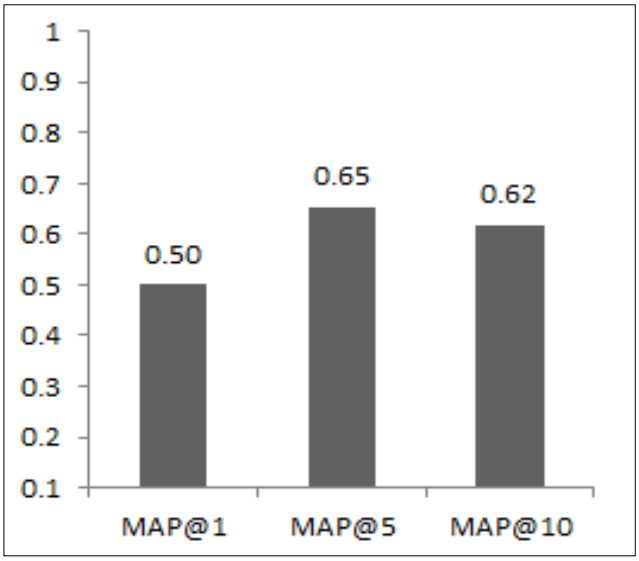


Figure 34: MAP Values for Topical Similarity Strategy

Figure 35 shows the percentages of the user disapproval reasons for Strategy 4. The most significant reason is "I am not interested in this user's tweets" which has 54%, the second popular one is "I can reach this information from other sources".

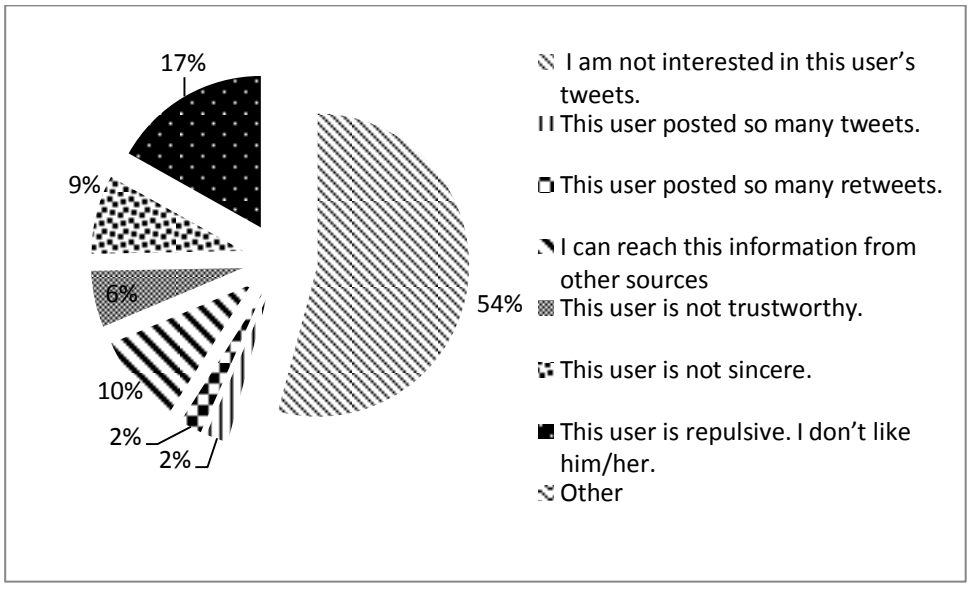


Figure 35: Disapproval Reasons for Topical Similarity Strategy

6.4.2.2 Opinion Similarity

In order to find the users that have similar taste, we combine the topical similarities with their sentiment values in this experiment. At the end of the experiment, users agree to follow 113 users out of 220 recommended users (51%). Figure 36 shows the mean average precision (MAP) values in @1, @5 and @10.

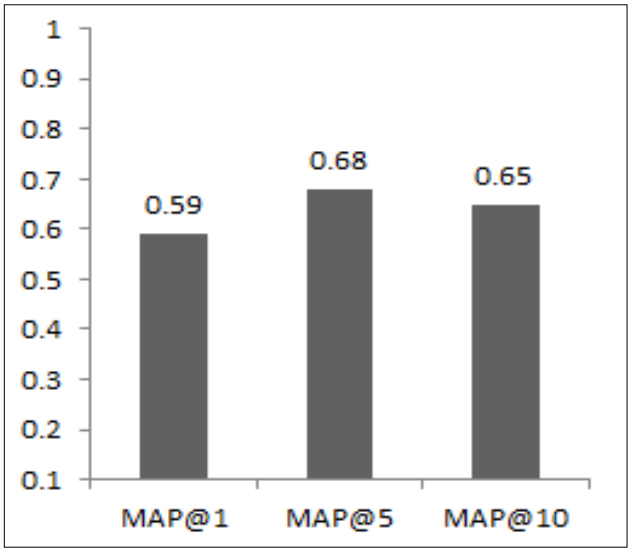


Figure 36: MAP Values for Opinion Similarity Strategy

The reasons of refusal by the users for Strategy 4 are shown in Figure 37. The most significant reason is "I am not interested in this user's tweets" which has 54%. The second popular one is "I can reach this information from other sources" with 14% percentage.

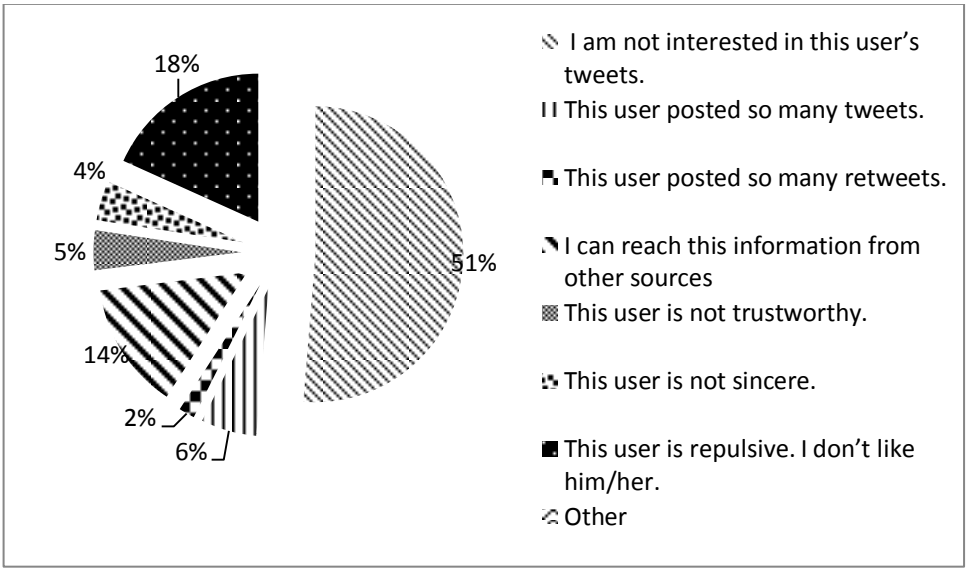


Figure 37: Disapproval Reasons for Opinion Similarity Strategy

6.4.2.3 Comparison of Content Analysis Strategies

Figure 38 shows the comparisons of the content based strategies' for MAP@1, MAP@5 and MAP@10. It is observed that the opinion similarity is only slightly

different from to topical similarity. As seen in Figure 38, in both strategies results are very similar to each other.

Recommended items in the ranked lists are 80% similar in both experiments. Since we ran topical and option strategy on the selected users, who are topologically closer to the user, similarity between results are expected.

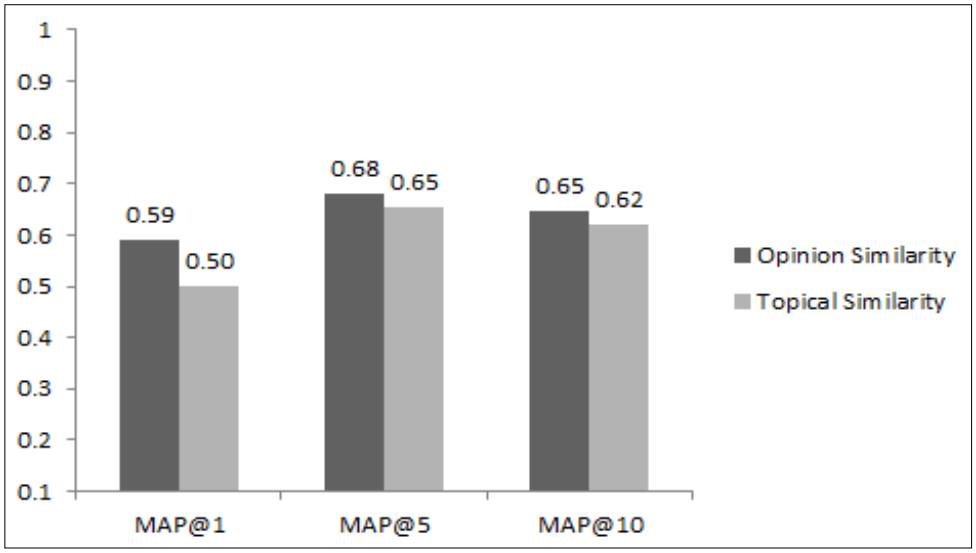


Figure 38: Comparison of MAP Values for Content-Base Strategies

6.4.2.4 Experiment on Combination of All Strategies

As stated before, combined strategy is formed by adding other strategies' normalized values. Figure 39 shows the performance of combined strategy for MAP@1, MAP@5 and MAP@10. In this experiment, out of 239 recommended items, 149 of them are approved to be followed by the survey participants.

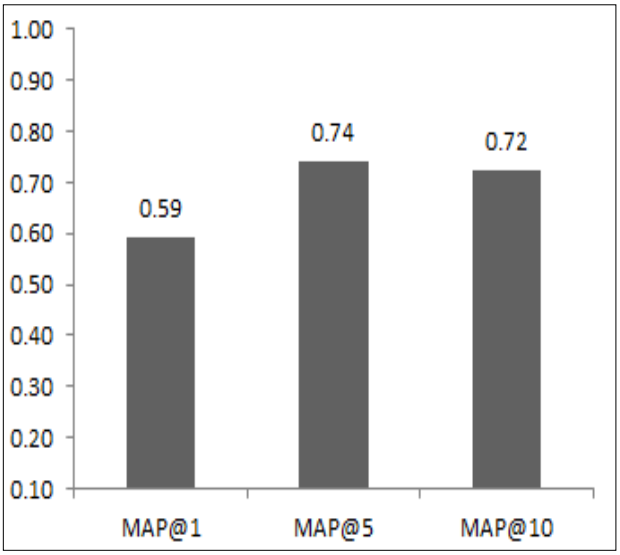


Figure 39: MAP Values for Combined Strategy

As seen in Figure 40, the most popular rejection reason is "I am not interested in this user's tweets" with 34%. The second reason is with 25%, "I can reach this information from other sources".

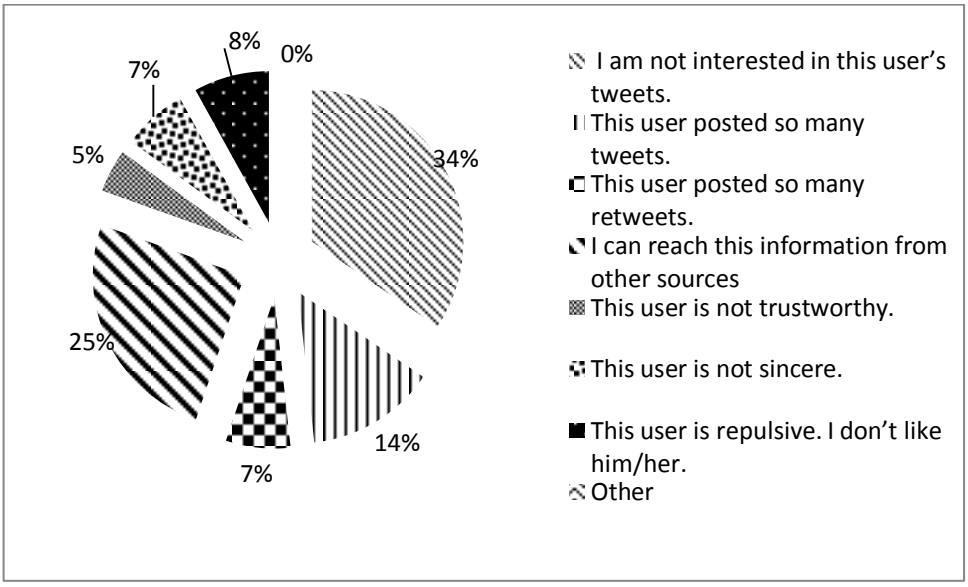


Figure 40: Disapproval Reasons for Combined Strategy

6.4.3 Comparison of All Strategies and Evaluation Results

In this section, all proposed strategies are evaluated. Firstly, we compare all the strategies at pMAP@1, MAP@5 and MAP@10 separately. Figure 41 shows the MAP@1 values for all the strategies. As seen in the Figure 41, followees of followees give the best result with 0.73 average precision value. The second best

result is the combination of topology-based approaches. The favorites of favorites has the lowest precision values at MAP@1 with 0.36.

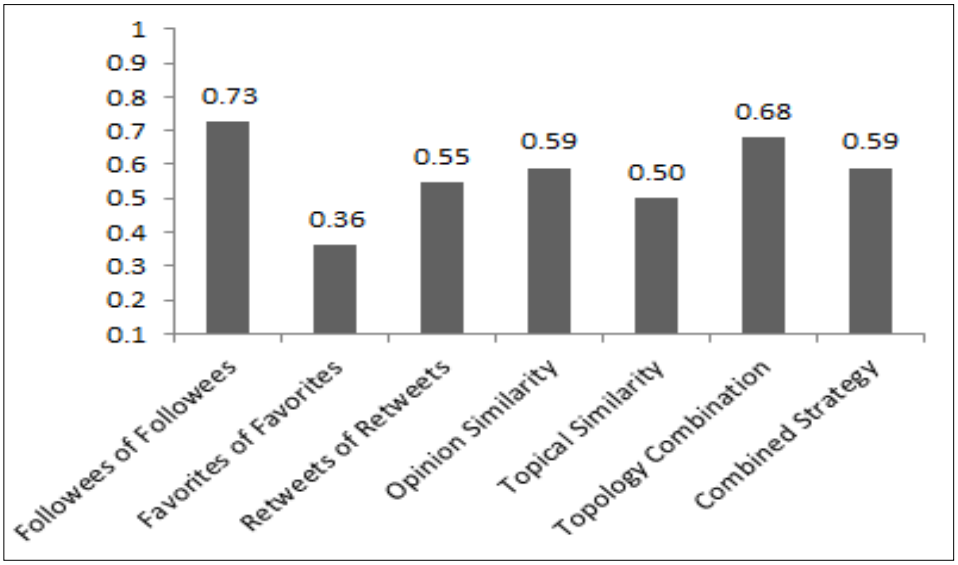


Figure 41: MAP@1 values

MAP@5 results are compared in Figure 42. Combination of topology works best among all strategies. The second best results are gained from combination of all strategies experiment. Favorites of favorites strategy gives the lowest results in comparison to other strategies.

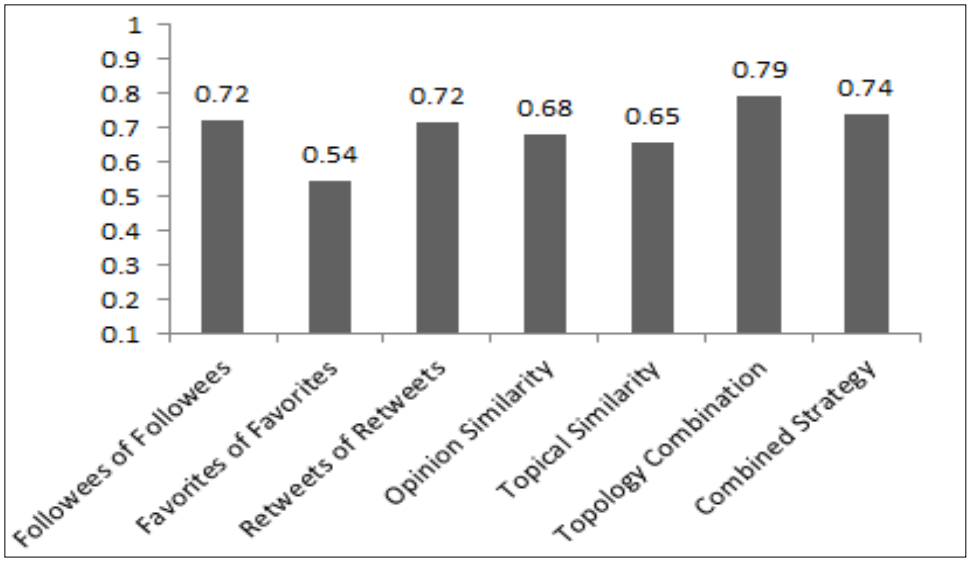


Figure 42: MAP@5 values

In Figure 43, the MAP@10 values are compared. As seen in the figure, topology combination performs better than the other strategies. The ordering of experiments in MAP@10 is similar to that of MAP@5. The second best one is combined strategy and favorites of favorites based strategy perform poorly.

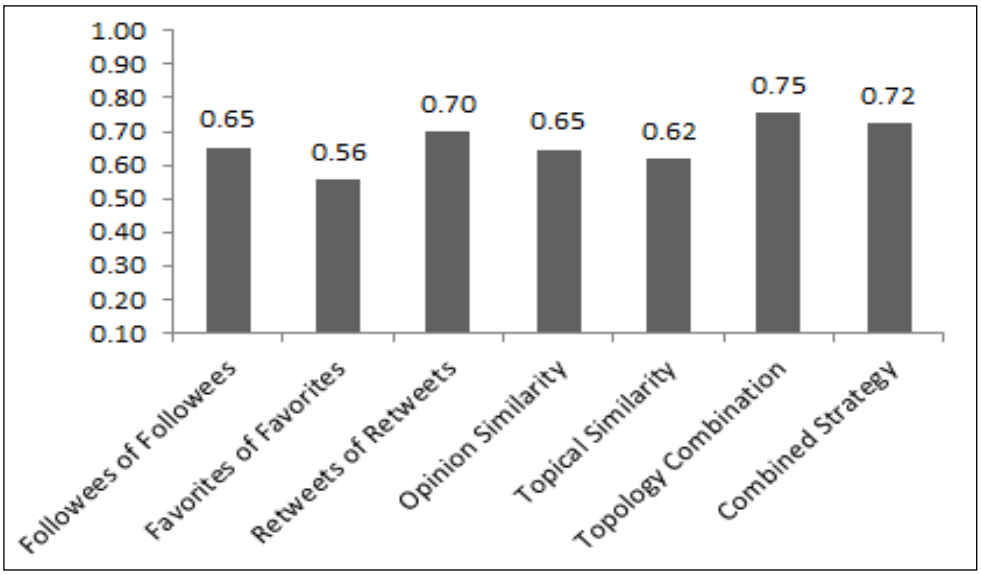


Figure 43: MAP@10

Figure 45 shows the comparison of MAP@1, MAP@5, and MAP@10 results. As seen in the Figure 44, the combined topological strategy (Strategy 6) generate better MAP score than other strategies and MAP@1 values are lower than MAP@5 and MAP@10 except Strategy1 (Followees of Followees).

We also found out that Favorites of Favorites (Strategy 2) have lowest performance value among all approaches. We can deduce that in user recommendation, retweets are more valuable data than favorites.

On the contrary to our expectations, content based analysis has negative effect on topological analysis. After integrating content analysis and combined topology strategies, as shown in Figure 45, the MAP values are slightly decreased.

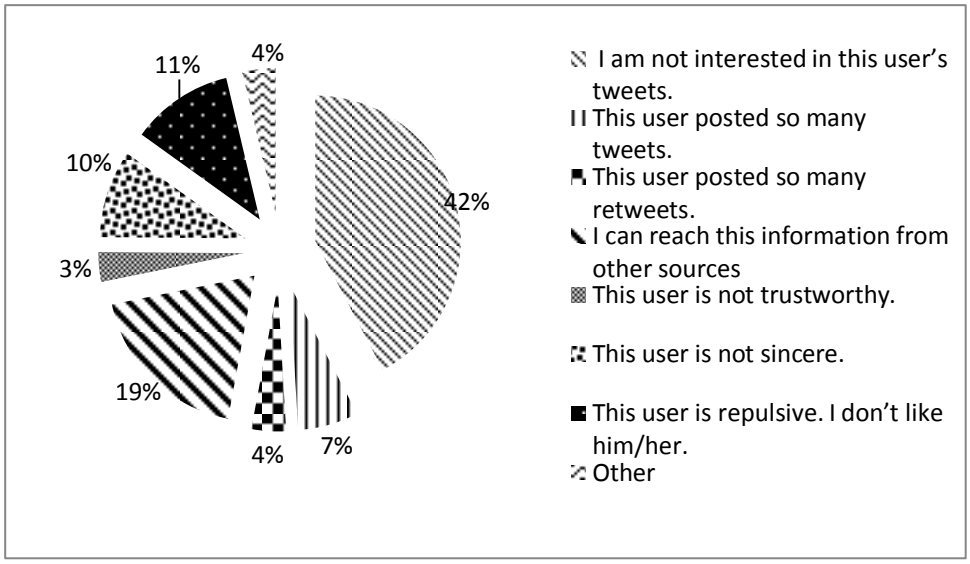


Figure 44: Disapproval Reasons

The disapproval reasons for all the strategies are shown in Figure 43. The most popular reason is "I am not interested in user's tweets" with 42% and "I can reach this information from other sources" takes the second place among other reasons. As mentioned before, when users reject to follow user, we show some reasons from a pop up window. Users are able to select more than one option or they can leave it empty. We observe that in the first strategy, users tend to give more details about why they do not want to follow recommended user. However, in the latest experiments, even we ask them to rate same user, they tend to give less detail. For example, similar users can be shown in recommendation lists. In the first three strategies, they tend to choose more than one solution. On the contrary, when the user comes across the same user in latest experiments, they do not want to give details about disapproval and they just clicked "I am not interested in user's tweets" option. Since finding not interesting to someone tweets is a more generic answer than others, users tend to choose it.

Finally, according to our experiments, it can be concluded that combination of topological approaches outperform the other strategies. We find that at MAP@5 values are higher than the other list lengths.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this thesis work, we propose a personalized followee recommendation for Twitter. Our aim is to help the active users for finding interesting people and overcome information overload problem. Our recommendations are mainly based on topological features of Twitter. Besides the following relationships, Twitter has different features, which are retweeting and favoriting. One of our aims in this study is finding the effects of these hidden features on followee gaining.

To begin with, we conduct a survey to understand Twitter users' retweet and favorite behaviours. Survey results show that people tend to favorite a tweet when it is funny or interesting. Additionally, people tend to retweet a tweet for broadcasting purpose. Users' favorite or retweet a tweet when they like it.

In this research, these features are used in order to make better followee suggestions to the users. In addition to the topological features, we combine topological methods with content- based analysis within the scope of English and Turkish language.

In this study, we elaborate on and compare seven different strategies in order to find the most effective way of recommending followees to the active Twitter users. We calculate the mean average precision (MAP) values under different ranking positions in order to compare our recommendation performance under different ranking positions. In the first three strategies, retweets, favorites and followee information were separately used to generate recommendations. The experiments show that recommendations that are based on user's retweets have better results among these three approaches for MAP@5 and MAP@10. In the fourth strategy, these three topological strategies normalized values are merged to enhance the effectiveness of recommendations. According to our experiments results, the merged strategy has shown better performance than other topological strategies. After finishing topology part, we conduct a user-centred personalized network topology by using users' followees, retweets and favorites. In the second part of the study, we include content based analysis to find more relevant users from user-centred topology. In content based approaches, our recommendations are based on topic similarity and opinion similarity between the users. In topical similarity, twitterLDA [13] is used for finding the topics from the tweets. In opinion analysis, in addition to the topical analysis, we also include user's sentiments. The experiments show that content-based approaches

perform worse than topology-based approaches. Lastly, in our last strategy, we merge all strategies. The experiments show that including content based approach decreases the relevancy of the recommended items.

In our recommender system, we show our users not only personally generated candidate lists but also we ask the reasons why they do not want to follow that recommended user. The given feedbacks from all strategies most common answer is "I am not interested in user's tweets" and "I can reach this information from other sources".

There are some limitations of this study. First of all, Twitter allows us to collect latest 3200 tweets. If a user publishes more than 3200 tweets, we are not able to collect after 3200 tweets due to the Twitter limitations[8]. Secondly, after data collection process, some accounts changed their privacy settings. These users were eliminated when constructing candidate lists. Thirdly, we believe that more participants may be included to our study. Lastly, some of our participants (user 6, user 16, and user 18) have less than 10 retweets and favorites, even though they have enough tweets. We will generate better recommendations if they have more retweets and favorites.

To conclude, we compare our proposed strategies and the combined topology algorithm performed the best while the worst performance is obtained in favorites of favorite's strategy.

The contributions of our research can be listed as following:

- We design a followee recommender system that compare and combine topological strategies and content-based strategies.
- We use users' favorite information separately while exploring new users in Twitter.
- We point out the differences between "retweet" and "favorite" features and compare the effectiveness of these features in followee recommendation in Twitter using survey results.
- We combine topical analysis with sentiment analysis on for both Turkish and English language.

[8] https://dev.twitter.com/rest/reference/get/statuses/user_timeline

7.2 Future Work

Finding relevant users in the social networks is a popular research topic. Many researchers come up with new ideas in this area every day. We would like to present some points that can be improved in the future.

- While rating the recommended users, collaborative parameters can be considered like popularity, retweet count or retweeted tweet count.
- Our experiments have shown that some users were not preferred to be followed because they generate too much content. This study can be expanded by considering the effects of tweet/retweet count and frequency.
- In this study we have counted the occurrences of seen users in the topology in order to rank them. This study can be expanded by using different ranking methods like PageRank or SimRank.

REFERENCES

1. Kywe, Su Mon, Ee-Peng Lim, and Feida Zhu. "A survey of recommender systems in twitter." *Social Informatics*. Springer Berlin Heidelberg, 2012. 420-433.
2. Hannon, John, Kevin McCarthy, and Barry Smyth. "Finding useful users on twitter: twittomender the followee recommender." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2011. 784-787.
3. Armentano, Marcelo G., Daniela L. Godoy, and Analía A. Amandi. "A topology-based approach for followees recommendation in Twitter." *Workshop chairs*. 2011.
4. Hutto, C. J., SaritaYardi, and Eric Gilbert. "A longitudinal study of follow predictors on twitter." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013.
5. Garcia, Ruth, and Xavier Amatriain. "Weighted content based methods for recommending connections in online social networks." *Workshop on Recommender Systems and the Social Web*. 2010.
6. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. 2010.
7. Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 2009.
8. Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007, May). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 171-180). ACM.
9. Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." *ICWSM* 11 (2011): 538-541.
10. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of Machine Learning Research* 3 (2003): 993-1022.
11. Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494). AUAI Press.
12. Akın, Ahmet Afsin, and Mehmet Dündar Akın. "Zemberek, an open source NLP framework for Turkic Languages." *Structure* 10 (2007).
13. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval* (pp. 338-349). Springer Berlin Heidelberg.
14. Hong, Liangjie, and Brian D. Davison. "Empirical study of topic modeling in twitter." *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010.
15. Erogul, U. "Sentiment analysis in Turkish." *Middle East Technical University, Ms Thesis, Computer Engineering* (2009).

16. Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish. In *Computer and Information Sciences III* (pp. 437-445). Springer London.Huberman,
17. Bernardo A., Daniel M. Romero, and Fang Wu. "Social networks that matter: Twitter under the microscope." *arXiv preprint arXiv: 0812.1045*(2008).
18. Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (socialcom), 2010 IEEE Second International Conference on*(pp. 177-184). IEEE.
19. "What are @replies and mentions?",https://support.twitter.com/articles/14023 , Last accessed: April 15, 2015.
20. Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 1031-1040). ACM.
21. Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th International Conference on World wide web (pp. 591-600). ACM.
22. Ricci, Francesco, LiorRokach, and BrachaShapira. *Introduction to recommender systems handbook*. Springer US, 2011.
23. Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.
24. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., ... & Sampath, D. (2010, September). The YouTube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 293-296). ACM.
25. "Google Prediction API ", https://cloud.google.com/prediction/docs/getting-started , Last accessed: April 15, 2015.
26. Schafer, J. Ben, Joseph Konstan, and John Riedl. "Recommender systems in e-commerce." *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 1999.
27. Vozalis, Emmanouil, and Konstantinos G. Margaritis. "Analysis of recommender systems algorithms." *Proceedings of the 6th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA-2003), Athens, Greece*. Vol. 2003. 2003.
28. Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291-324). Springer Berlin Heidelberg.
29. Guy, I., Zwerdling, N., Ronen, I., Carmel, D., & Uziel, E. (2010, July). Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 194-201). ACM.
30. Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." *Advances in Artificial Intelligence* 2009 (2009): 4.
31. Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. (2013, May). Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 505-514). International World Wide Web Conferences Steering Committee.

32. Kamath, K., Sharma, A., Wang, D., & Yin, Z. (2014). RealGraph: User Interaction Prediction at Twitter. In *User Engagement Optimization Workshop@ KDD*.
33. Goel, A., Sharma, A., Wang, D., & Yin, Z. (2013). Discovering Similar Users on Twitter. In *11th Workshop on Mining and Learning with Graphs*.
34. Balasubramanyan, Ramnath, and Aleksander Kolcz. "“w00t! feeling great today!” Chatter in Twitter: Identification and Prevalence." *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 2013.
35. Yang, S. H., Kolcz, A., Schlaikjer, A., & Gupta, P. (2014, August). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (pp. 1907-1916). ACM.
36. Lin, Jimmy, and Alek Kolcz. "Large-scale machine learning at twitter."*Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012.
37. Burke, Robin. "Knowledge-based recommender systems." *Encyclopedia of Library and Information Systems* 69.Supplement 32 (2000): 175-186.
38. Burke, Robin. "Hybrid recommender systems: Survey and experiments." *User Modeling and User-Adapted Interaction* 12.4 (2002): 331-370.
39. Oflazer, Kemal. "Turkish and its challenges for language processing."*Language Resources and Evaluation*: 1-15.
40. Kumar, Ravi, Jasmine Novak, and Andrew Tomkins. "Structure and evolution of online social networks." *Link mining: models, algorithms, and applications*. Springer New York, 2010. 337-357.
41. O'Reilly, Tim, and Sarah Milstein. *The twitter book*. “O’Reilly Media, Inc.", 2011.
42. Tavakolifard, Mozhgan, Kevin C. Almeroth, and Jon Atle Gulla. "*Does social contact matter?: modelling the hidden web of trust underlying twitter.*"Proceedings of the 22nd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2013.
43. McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in social networks." *Annual review of sociology* (2001): 415-444.
44. Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 261-270). ACM.
45. Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. "*Characterizing Microblogs with Topic Model*s." ICWSM 10 (2010): 1-1.
46. Kaya, Mesut, Guven Fidan, and Ismail H. Toroslu. "*Sentiment analysis of turkish political news."* Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2012.
47. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.
48. Dann, Stephen. "*Twitter content classification."* First Monday 15.12 (2010).

49. "Twitter4J", http://twitter4j.org/en/, Last accessed: April 15, 2015.
50. Jennifer. "*Combining provenance with trust in social networks for semantic web content filtering*." Provenance and Annotation of Data. Springer Berlin Heidelberg, 2006. 101-108.
51. Bhuiyan, Touhid. "*A survey on the relationship between trust and interest similarity in online social networks*." Journal of Emerging Technologies in Web Intelligence 2.4 (2010): 291-299.
52. Hussain, Farookh Khadeer, Omar Khadeer Hussain, and Elizabeth Chang. *"An overview of the interpretations of trust and reputation."* Emerging Technologies and Factory Automation, 2007. ETFA. IEEE Conference on. IEEE, 2007.
53. D Harrison McKnight, Norman L. Chervany. *"What trust means in e-commerce customer relationships: an interdisciplinary conceptual typology."* International Journal of Electronic Commerce 6.2 (2001): 35-59.
54. Jeh, Glen, and Jennifer Widom. "*SimRank: a measure of structural-context similarity.*" Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002.
55. Golder, S. A., Yardi, S., Marwick, A., & Boyd, D. (2009). A structural approach to contact recommendations in online social networks. In *Workshop on Search in Social Media, SSM*.
56. Celebi, H. Burak, and Susan Uskudarli. "*Content Based Microblogger Recommendation*." Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). IEEE, 2012.
57. Gemci, Fahriye, and Kadir A. Peker. "*Extracting Turkish tweet topics using LDA.*" Electrical and Electronics Engineering (ELECO), 2013 8th International Conference on. IEEE, 2013.

APPENDIXES

Appendix A: SURVEY

A.1. Survey Questions (in Turkish)

Twitter' da neden Favori- RT kullanırsınız?

Bu anket, Twitterda kullanicilarin hangi sebeblerle favori/RT kullandiklarini ogrenmek icin bir tez çalışması kapsamında hazırlanmıştır. Anketimiz aktif twitter kullanıcıları içindir. Yanıtlarınız sadece akademik çalışma amaçlı kullanılacaktır. Sorulara vereceğiniz samimi yanıtlar mevcut durumun ortaya konması açısından önem taşımaktadır. Şimdiden zamanınızı ayırarak çalışmaya sağladığınız katkı için çok teşekkür ederim.

AysuYanar
aysudagli86@gmail.com

1. **Twitter Kullanıcı Adınızı Yazar mısınız? (Gizli kalacaktır)**

2. **Twitter' da kaç kişiyi takip ediyorsunuz?**
   - <10
   - 10-100
   - 100-1000
   - 1000-10000
   - 10000-100000
   - 100000<

3. **Twitter' da kaç takipçiniz var?**
   - <10
   - 10-100
   - 100-1000
   - 1000-10000
   - 10000-100000
   - 100000<

4. **Twitter' ı hangi sıklıkla kullanıyorsunuz?**
   - Ayda yılda bir
   - Haftada bir
   - Günde bir
   - Günde birçok kez

5. **Twitter kullanma nedeniz nedir? (bir veya daha çok şıkkı işaretleyebilirsiniz)**
   - Haber almak için
   - Eğlenmek için
   - Arkadaşlarımla iletişim kurmak için
   - Kendim hakkında haber, bilgi vermek için
   - Ünlüleri takip etmek için
   - Vakit geçirmek için

6. **Ne zamandır Twitter kullanıyorsunuz?**
   - >1 yıl
   - Ay - 1 yıl
   - 1 ay - 6 ay
   - 1 ay
   - Yeni üyeyim

7. **Twitter' daki "fav" ile Facebook' taki "like" birbirine benziyor diyebilir miyiz?**

- Hiç Katılmıyorum
- Katılmıyorum
- Kararsızım
- Katılıyorum
- Kesinlikle Katılıyorum

8. Bir Tweet i neden Favorite(fav) edersiniz?

	Hiç Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
1) Sonradan okumak için					
2) Komik olduğu için					
3) Patronumu - is arkadaşımı (saygıdan)					
4) Arkadaşımın paylaşımına destek olmak için					
5) Beğendiğim için					
6) Cok ilginç bulduğum zaman					
7) Flört etmek için					
8) Onayladığımı göstermek için					
9) Teşekkür etmek için					
10) Nefret ettiğim zaman					

9)Tweet i neden Retweet(RT) edersiniz

	Hiç Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
1) Haberi/Tweet yaymak için					
2) Alıntı yapmak için					
3) Tesekkur etmek için					
4) Beğendiğim zaman					
5) Nefret ettiğim/ Karşı olduğum bir bilgiyi göstermek için					
6) Arkadaşıma destek olmak için					
7) Komik bulduğum zaman					
8) Takipçi edinmek için					
9) Reklam için					
10) Sonradan okumak için					

A.1. Survey Answers (in Turkish)

1. Twitter' da kaç kişiyi takip ediyorsunuz?

Answer Choices	Responses
%	0.00%
<10	0
%	27.27%
10-100	18
%	66.67%
100-1000	44
%	6.06%
1000-10000	4
%	0.00%
10000-100000	0
%	0.00%
100000<	0
Total	**66**

2. Twitter' da kaç takipçiniz var?

Answer Choices	Responses
%	1.52%
<10	1
%	28.79%
10-100	19
%	57.58%
100-1000	38
%	12.12%
1000-10000	8
%	0.00%
10000-100000	0
%	0.00%
100000<	0
Total	**66**

3. Twitter' ı hangi sıklıkla kullanıyorsunuz?

Answer Choices	Responses
%	4.55%
Ayda yılda bir	3
%	9.09%
Haftada bir	6
%	24.24%
Günde bir	16
%	62.12%
Günde birçok kez	41
Total	**66**

4. Twitter kullanma nedeniz nedir? (bir veya daha çok şıkkı işaretleyebilirsiniz)

Answer Choices	Responses
%	92.42%
Haber almak için	61
%	53.03%
Eğlenmek için	35
%	33.33%
Arkadaşlarıml a iletişim kurmak için	22
%	30.30%
Kendim hakkında haber, bilgi vermek için	20
%	12.12%
Ünlüleri takip etmek için	8
%	57.58%
Vakit geçirmek için	38
Total Respondents: 66	

5. Ne zamandır Twitter kullanıyorsunuz?

Answer Choices	Responses
%	96.92%
>1 yıl	63
%	1.54%
6 ay - 1 yıl	1
%	1.54%
1 ay - 6 ay	1
%	0.00%
1 ay	0
%	0.00%
yeni üyeyim	0
Total	**65**

6. Twitter' daki "fav" ile Facebook' taki "like" birbirine benziyor diyebilir miyiz?

	Hiç Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum	Total	Weighted Average
%	3.03%	27.27%	15.15%	40.91%	13.64%		
1	2	18	10	27	9	66	3,35

7. Bir Tweet'i Neden Favorite(fav) Edersiniz?

	Hiç Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum	Total	Weighted Average
%	11.29%	9.68%	3.23%	50.00%	25.81%		
1) Sonradan okumak için	7	6	2	31	16	**62**	**3,69**
%	3.23%	4.84%	9.68%	56.45%	25.81%		
2) Komik olduğu için	2	3	6	35	16	**62**	**3,97**
%	45.00%	38.33%	10.00%	5.00%	1.67%		
3) Patronumu - is arkadaşımı (saygıdan)	27	23	6	3	1	**60**	**1,8**
%	13.33%	15.00%	18.33%	45.00%	8.33%		
4) Arkadaşımın paylaşımına destek olmak için	8	9	11	27	5	**60**	**3,2**
%	1.54%	1.54%	0.00%	44.62%	52.31%		
5) Beğendiğim için	1	1	0	29	34	**65**	**4,45**
%	0.00%	3.28%	6.56%	52.46%	37.70%		
6) Cok ilginç bulduğum zaman	0	2	4	32	23	**61**	**4,25**
%	50.85%	28.81%	5.08%	11.86%	3.39%		
7) Flört etmek için	30	17	3	7	2	**59**	**1,88**
%	4.69%	9.38%	6.25%	53.13%	26.56%		
8) Onayladığımı göstermek için	3	6	4	34	17	**64**	**3,88**
%	14.75%	19.67%	14.75%	34.43%	16.39%		
9) Tesekkur etmek için	9	12	9	21	10	**61**	**3,18**
%	83.05%	13.56%	1.69%	1.69%	0.00%		
10) Nefret ettiğim zaman	49	8	1	1	0	**59**	**1,22**

8. Bir Tweeti Neden Retweet(RT) Edersiniz?

	Hiç Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum	Total	Weighted Average
%	0.00%	3.08%	0.00%	30.77%	66.15%		
1) Haberi/ Tweet' i yaymak için	0.0	2.0	0.0	20.0	43.0	**65**	**4,60**
%	3.28%	6.56%	4.92%	44.26%	40.98%		
2) Alıntı yapmak için	2.0	4.0	3.0	27.0	25.0	**61**	**4,13**
%	25.86%	31.03%	18.97%	17.24%	6.90%		
3) Tesekkur etmek için	15.0	18.0	11.0	10.0	4.0	**58**	**2,48**
%	0.00%	1.54%	9.23%	52.31%	36.92%		
4) Beğendiğim zaman	0.0	1.0	6.0	34.0	24.0	**65**	**4,25**
%	27.87%	22.95%	16.39%	21.31%	11.48%		
5) Nefret ettiğim/ Karşı olduğum bir bilgiyi göstermek için	17.0	14.0	10.0	13.0	7.0	**61**	**2,66**
%	8.20%	11.48%	13.11%	49.18%	18.03%		
6) Arkadaşıma destek olmak için	5.0	7.0	8.0	30.0	11.0	**61**	**3,57**
%	6.35%	7.94%	7.94%	46.03%	31.75%		
7) Komik bulduğum zaman	4.0	5.0	5.0	29.0	20.0	**63**	**3,89**
%	69.49%	20.34%	6.78%	3.39%	0.00%		
8) Takipçi edinmek için	41.0	12.0	4.0	2.0	0.0	**59**	**1,44**
%	60.00%	25.00%	3.33%	8.33%	3.33%		
9) Reklam için	36.0	15.0	2.0	5.0	2.0	**60**	**1,70**
%	42.37%	22.03%	13.56%	18.64%	3.39%		
10) Sonradan okumak için	25.0	13.0	8.0	11.0	2.0	**59**	**2,19**

Appendix B: STOP WORD LISTS

B.1 Turkish Stop Word List

sadece	yi	altmıs	milyar
hep	le	altı	milyon
olur	ye	bana	mu
te	ta	ben	mı
nda	ten	benden	mü
mış	tan	beni	nasıl
bb	ş	benim	ne
ıı	da	bey	nerde
sonra	de	bin	nerede
bazen	ise	bir	nereye
ilk	için	biri	o
bile	daha	birini	olan
olmak	icin	biz	olarak
bugün	kendi	bizde	olsa
ederim	2	bizi	olup
simdi	3	bize	olursa
şimdi	4	bizden	on
yok	5	bizi	ona
zaten	6	bizim	ondan
http	7	bu	onlar
https	8	buna	onlardan
çok	9	bunda	onları
değil	ç	bunlar	onların
yla	ğ	bunları	onu
şey	ö	bunların	onun
ve	ş	bunu	otuz
ki	ü	bunun	öyle
la	ı	burada	pek
nın	İ	da	rt
nin	"	de	sekiz
te	":	doksan	seksen

de	via	dokuz	sen
değil	http://t.co	dolayısıyla	senden
degil	yedi	dört	seni
diye	yetmiş	edecek	senin
cok	yirmi	eden	siz
bi	yüz	ederek	sizden
mi	ve	edilecek	sizi
mı	veya	ediliyor	sizin
ın	ya	edilmesi	şey
kadar	yani	ediyor	şeyden
dan	yoksa	elli	şeyi
den	trilyon	en	şeyler
den	tüm	etmesi	şöyle
itibariyle	üç	etti	şu
katrilyon	üzere	ettiği	şuna
ki	var	ettiğini	şunda
kim	vardı	gibi	şundan
kimden	ama	hangi	şunları
kime	fakat	herhangi	şunu
kimi	lakin	iki	ilgili
kırk	pe	ile	işte
falan	hala		
	ise		

B.2 English Stop Word List

c	li	i	o
c'mon	el	i'd	obviously
c's	es	i'll	of
came	los	i'm	off
can	est	i've	often
can't	lo	ie	oh
cannot	tu	if	ok
cant	les	ignored	okay
cause	con	immediate	old
causes	su	in	on
certain	se	inasmuch	once
certainly	del	inc	one
changes	di	indeed	ones
clearly	je	indicate	only
co	em	indicated	onto
com	una	indicates	or
come	don	inner	other
comes	min	insofar	others
concerning	mins	instead	otherwise
consequently	a	into	ought
consider	a's	inward	our
considering	able	is	ours
contain	about	isn't	ourselves
containing	above	it	out
contains	according	it'd	outside
corresponding	accordingly	it'll	over
could	across	it's	overall
couldn't	actually	its	own
course	after	itself	p
currently	afterwards	j	particular
d	again	just	particularly
definitely	against	k	per
described	ain't	keep	perhaps
despite	all	keeps	placed
did	allow	kept	please
didn't	allows	know	plus
different	almost	knows	possible
do	alone	known	presumably
does	along	l	probably
doesn't	already	last	provides
doing	also	lately	q
don't	although	later	que

done	always	latter	quite
down	am	latterly	qv
downwards	among	least	r
during	amongst	less	rather
e	an	lest	rd
each	and	let	re
edu	another	let's	really
eg	any	like	reasonably
eight	anybody	liked	regarding
either	anyhow	likely	regardless
else	anyone	little	regards
elsewhere	anything	look	relatively
enough	anyway	looking	respectively
entirely	anyways	looks	right
especially	anywhere	ltd	s
et	apart	m	said
etc	appear	mainly	same
even	appreciate	many	saw
ever	appropriate	may	say
every	are	maybe	saying
everybody	aren't	me	says
everyone	around	mean	second
everything	as	meanwhile	secondly
everywhere	aside	merely	see
ex	ask	might	seeing
exactly	asking	more	seem
example	associated	moreover	seemed
except	at	most	seeming
f	available	mostly	seems
far	away	much	seen
few	awfully	must	self
fifth	b	my	selves
first	be	myself	sensible
five	became	n	sent
followed	because	name	serious
following	become	namely	seriously
follows	becomes	nd	seven
for	becoming	near	several
former	been	nearly	shall
formerly	before	necessary	she
forth	beforehand	need	should
four	behind	needs	shouldn't
from	being	neither	since
further	believe	never	six
furthermore	below	nevertheless	so
g	beside	new	some
get	besides	next	somebody
gets	best	nine	somehow
getting	better	no	someone
given	between	nobody	something

gives	beyond	non	sometime
go	both	none	sometimes
goes	brief	noone	somewhat
going	but	nor	somewhere
gone	by	normally	soon
got	uses	not	sorry
gotten	using	nothing	specified
greetings	usually	novel	specify
h	uucp	now	specifying
had	which	nowhere	still
hadn't	while	three	sub
happens	whither	through	such
hardly	who	throughout	sup
has	who's	thru	sure
hasn't	whoever	thus	t
have	whole	to	t's
haven't	whom	together	take
having	whose	too	taken
he	why	took	tell
he's	will	toward	tends
hello	willing	towards	th
help	wish	tried	than
hence	with	tries	thank
her	within	truly	thanks
here	without	try	thanx
here's	won't	trying	that
hereafter	wonder	twice	that's
hereby	would	two	thats
herein	would	v	the
hereupon	wouldn't	value	their
hers	x	various	theirs
herself	y	very	them
hi	yes	via	themselves
him	yet	viz	then
himself	you	vs	thence
his	you'd	w	there
hither	you'll	want	there's
hopefully	you're	wants	thereafter
how	you've	was	thereby
howbeit	your	wasn't	therefore
however	yours	way	therein
whence	yourself	we	theres
whenever	yourselves	we'd	thereupon
where	z	we'll	these
where's	zero	we're	they
whereafter	think	we've	they'd
whereas	third	welcome	they'll
whereby	this	well	they're
wherein	thorough	went	they've
whereupon	thoroughly	were	whatever

wherever whether	those though	weren't what	when what's

TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☑

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı: Dağlı……

Adı: Aysu……

Bölümü: Bilişim Sistemleri ……...……

TEZİN ADI (İngilizce) : COMBINING TOPOLOGY-BASED & CONTENT-BASED ANALYSIS FOR FOLLOWEE RECOMMENDATION ON TWITTER

TEZİN TÜRÜ : Yüksek Lisans ☑ Doktora ☐

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. ☑
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden ☐
   kaynak gösterilmek şartıyla fotokopi alınabilir.
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. ☐

Yazarın imzası ..... Tarih .....