IDENTIFYING COMMUNITIES USING COLLABORATION AND WORD ASSOCIATION NETWORKS IN TURKISH SOCIAL MEDIA

A Thesis Submitted to

the Graduate School of Engineering and Sciences of

İzmir Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Abdullah Asil ATAY

December 2018
İZMİR

We approve the thesis of **Abdullah Asil ATAY**

Examining Committee Members:

_____

Assist. Prof. Dr. Selma TEKİR
Department of Computer Engineering, İzmir Institute of Technology

_____

Assoc. Prof. Dr. Tuğkan TUĞLULAR
Department of Computer Engineering, İzmir Institute of Technology

_____

Assist. Prof. Dr. Mutlu BEYAZIT
Department of Computer Engineering, Yaşar University

27 December 2018

_____

Assist. Prof. Dr. Selma TEKİR
Supervisor, Department of Computer Engineering, İzmir Institute of Technology

_____

Assoc. Prof. Dr. Tolga AYAV
Head of Department of Computer Engineering

_____

Prof. Dr. Aysun SOFUOĞLU
Dean of Graduate School of Engineering and Science

ACKNOWLEDGMENTS

First, I would like to express sincere appreciation to my supervisor Assist. Prof Dr. Selma TEKİR. I have learned a lot from with her special perspective and knowledge. It has been very exciting to work on this thesis.

I would like to express my thanks to my employer Yaşar Bilgi Company, my manager Volkan Abur and my valuable team for their unlimited support and encouragement during my thesis period. Their brilliant ideas were very important for me.

I would also like to express my special thanks to my dearest friend Alpay Özsüer from high school and mechanical engineering master science program and express my special thanks to Oktay Doğaner from work. They always motivated and supported me during my thesis period.

Lastly, I would like to express my special thanks to my all family members. Their encouragement is invaluable for me.

ABSTRACT

IDENTIFYING COMMUNITIES USING COLLABORATION AND WORD ASSOCIATION NETWORKS IN TURKISH SOCIAL MEDIA

Social media contents are always very attractive title for researchers. Scores of people use social media and share their ideas with pictures, videos or documents. Researchers analyze this information and they try to deduce beneficial data. A lot of researchers think that analyzing social media information is a very important research area.

There are a lot of social media platforms which have Turkish contents. We can give an example Ekşisözlük which have Turkish contents and popular social media platform in Turkey. Within scope of the thesis, Ekşisözlük contents downloaded, decomposed and used actively.

Social media consists of human or human made products and sharing contents have some similarities. In this thesis, to calculate similarities, some methods are used. Scope of the thesis, two different networks are created from same content which are word association network and collaboration network.

Word association network is a network that created by coexistence of words in specific window size. Collaboration network is a network that created by entered content to same title with different users. This information gives the similarity of users. These two networks are analyzed separately and deduced some information.

ÖZET

TÜRKÇE SOSYAL MEDYA'DAKİ İŞBİRLİĞİ VE SÖZCÜK BAĞLANTISI AĞLARINI KULLANARAK TOPLULUKLARIN BELİRLENMESİ

Sosyal medya, bir çok araştırma başlığı için hep ilgi çekici bir içerik olmuştur. Dünya üzerindeki bir çok insan sosyal medyayı kullanır ve çeşitli platformlarda fikirlerini paylaşır, düşüncelerini kelimelere, resimlere veya videolara döker. Paylaşılan bu bilgileri incelemek ve çıkarımlar yapmak, bir çok araştırmacı tarafından önemli bir araştırma alanı olarak görülmektedir.

İnternet ortamında bir çok ana dili Türkçe olan sosyal medya platformu vardır. Bunlardan bir tanesi de Ekşisözlük' tür. Bu tez kapsamında Ekşisözlük sosyal medya ortamından çekilen veriyle bir veritabanı oluşturulmuş ve ayrıştırılan bu veri araştırma süreçlerinde aktif olarak kullanılmıştır.

Bu tez kapsamında kelimelerden oluşan büyük, gerçek bir sosyal medya verisi incelenmiş, çeşitli benzerlik teknikler kullanılarak anlamlı sonuçlar elde edilmeye çalışılmıştır. Tez çalışmasında kullanılan sosyal medya verisinin incelenmesi ve anlamlaştırılması sırasında aynı veriden iki farklı ağ yapısı üretilmiştir. Bu ağ yapıları; Sözcük bağlantılı ağlar ve iş birliği ağlarıdır. Oluşturulan iki farklı ağ ayrı ayrı incelenip yorumlanarak, bazı matematiksel temellere dayanan çıkarımlar yapılmıştır.

Sözcük bağlantılı ağ yapısı, genel olarak kelimelerin birbiriyle kullanım sıklığına bakılarak oluşturulmuş bir ağdır. İş birliği ağları ise, Ekşisözlük ortamında çeşitli başlıklara yorum giren kullanıcıların ortak olarak kaç tane başlığa yorum girdiği sorusundan yola çıkılarak oluşturulmuş farklı bir ağ yapısıdır. İki farklı ağ yapısı incelenerek, bu tez kapsamında çeşitli sonuçlar üretilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

As social media sites continue to grow in popularity, people leave a mark from their characters with their shared objects. Many people exchange ideas, feelings, personal information in social media environment. Social media is a rich information source and social network analysis can reveal interesting relationships among people in networks.

People can be categorized by their shares on social media and these categories may be called as communities. A community is a small or large social unit who have something common like norms, religion values or identity. In formal terms, a community (also referred to as a module or a cluster) is typically thought of as a group of nodes with more connections amongst its members than between its members and the remainder of the network [28]. In general, nodes represent individual users/actors/items/resources whereas edges represent the link/flow of interaction/relationship among users. Nodes are building blocks for community. Also, communities are building blocks for networks. Combination of many communities create the network and social media environment has a lot of communities.

A network is said to have community structure if the nodes of the network can easily be grouped using some common attributes. Networks constructed out of social media environment are difficult to analyze due to complex relationships and data oversize. Because of the complexity, networks can be separated into communities and communities can be analyzed and researchers can get a result about network. Communities are smaller units than the whole network that's why analysis of the communities is easier than network.

Community detection is an important task in the study of networks as it provides information about the overall network structure in depth. Community detection algorithms are used to identify communities and their hierarchical organization in graphs. Community detection algorithms can be used in many different areas. Finding a common research area in collaboration networks (citation network), web page

clustering, finding a set of like-minded users for marketing and recommendations, prediction of missing links and the identification of false links in networks, finding protein interaction networks in biological networks and musical rhythmic pattern extraction may be given as examples.

Social media contents are artifacts. In real social media networks, nodes can be a member of more than one community in a network and thus communities overlap [28]. For example, people may share the same hobbies in social media or mechanical engineering and computer engineering students may take some common courses to graduate. Overlapping communities can be identified by overlapping community detection algorithms.

Overlapping community detection algorithms have different perspectives. In overlapping community detection logic, communities have sparse overlaps or dense overlaps, but communities are assumed to have no overlap in traditional community detection algorithms. At first, these algorithms are designed to work with sparse overlaps. In the context of network science, a sparse network is a network with a smaller number of links than the maximum possible number of links within the same network. Networks which have sparse overlapping communities are less densely connected than the non-overlapping parts of communities. Dense networks are much more applicable for real social networks because of human factor. In dense networks, the number of shared nodes increases and network where communities overlap tends to be more densely connected. State-of-the-art algorithms for overlapping community detection are designed to find out dense overlaps and are applicable to social media research. Overlapping community detection algorithms can be applied to word association and collaboration networks to gain insights about them.

Word association network is a network of words where words have relationships with each other in terms of semantic similarity. If a word has a relationship with another word and there are so many relationships like that, all words and their relationships combine and create communities. Then communities are combined, and they create the network. In general, the semantic similarity values are calculated with respect to the co-occurrence of words within a fixed-size window all along a given corpus.

Collaboration is a joint effort of multiple individuals or work groups to accomplish a task or a project. A collaboration network consists of components which try to help each other. For instance, cloud computing is a good example for a

collaboration network. Computers are nodes and they connect with each other. These connections (relationships) and nodes create the collaboration network. In social networks, social media users collaborate to each other with their shared objects. Objects can be text, photos or videos.

In this study, we have real, Turkish social media data from a social media website. We construct a word association network and a collaboration network from the same corpus. Our aim is to find interesting patterns from these networks and match some attributes between them. For instance, could similar users, which post objects into similar topics, use similar words? In analyzing these networks, state-of-the-art overlapping community detection algorithms are used as they are appropriate to the overlapping nature of the data.

Before the overlapping community detection algorithms, social media data are analyzed with community detection algorithms. In these algorithms, detected communities do not share any nodes. But communities which are created with overlapping community detection algorithms have shared nodes. In large networks, community detection algorithms cannot show communities correctly and they have trouble about it [28]. In this study, we have large social media data and we compare these two perspectives experimentally. Evidence shows that the detected communities using overlapping and non-overlapping community detection algorithms vary in a statistically significant way.

After the analysis, we reach some conclusions. Firstly, a word association network and a collaboration network are created from the same source but using different types of data. We found some connections between these two different networks. Secondly, the application of overlapping and non-overlapping community detection algorithms create different clusterings on our data. Then the results show that we can see similar users in social collaboration network use similar words. This research is done on Turkish social media and contributes to Turkish language processing and resources.

Chapter 1 presents the purpose of the thesis; problem definition issues and shows the challenges. Chapter 2 gives background information. Chapter 3 explains related work. Chapter 4 presents experimental operations. Data collection process, natural language preprocessing operations, creation and analysis of collaboration and word association networks, application of overlapping community detection algorithms

and comparison of overlapping and non-overlapping structures will show up. At the end, thesis concludes with Chapter 5, which discusses conclusions of this study.

CHAPTER 2

BACKGROUND

This chapter presents background information of the thesis. Community structure, word association and collaboration network structures and state-of-the-art overlapping community detection algorithms will be presented. At the end of this chapter, overlapping community detection algorithm selection will be explained.

2.1. Network

Network is a group of elements or nodes which communicate with each other through edges. Networks provide a powerful framework to model real world relationships in which nodes denote entities, and links indicate the interactions between these entities [16]. In social network analysis, for instance, nodes can represent people and links can represent friendship or shared objects between individuals [16].

2.1.1. Network Types

In social network analysis, different network types are used. In this part, word association network and collaboration network will be explained.

2.1.1.1. Word Association Network

Word association network is a network of words where words have relationships with each other in terms of semantic similarity. An example for a word association network is depicted in Figure 2.1. In Figure 2.1, there is a word association network that has five communities and their connections. Feeling of lightness, colors / weather, visible light, electric light and industrial light are the names of communities. The light

word has many relationships with other words that's why it is a sharing word for all communities and it is on the center of the network. Word association networks are created from big text corpora and analyzed to provide insight with respect to the syntactic and semantic understanding of the data easily.

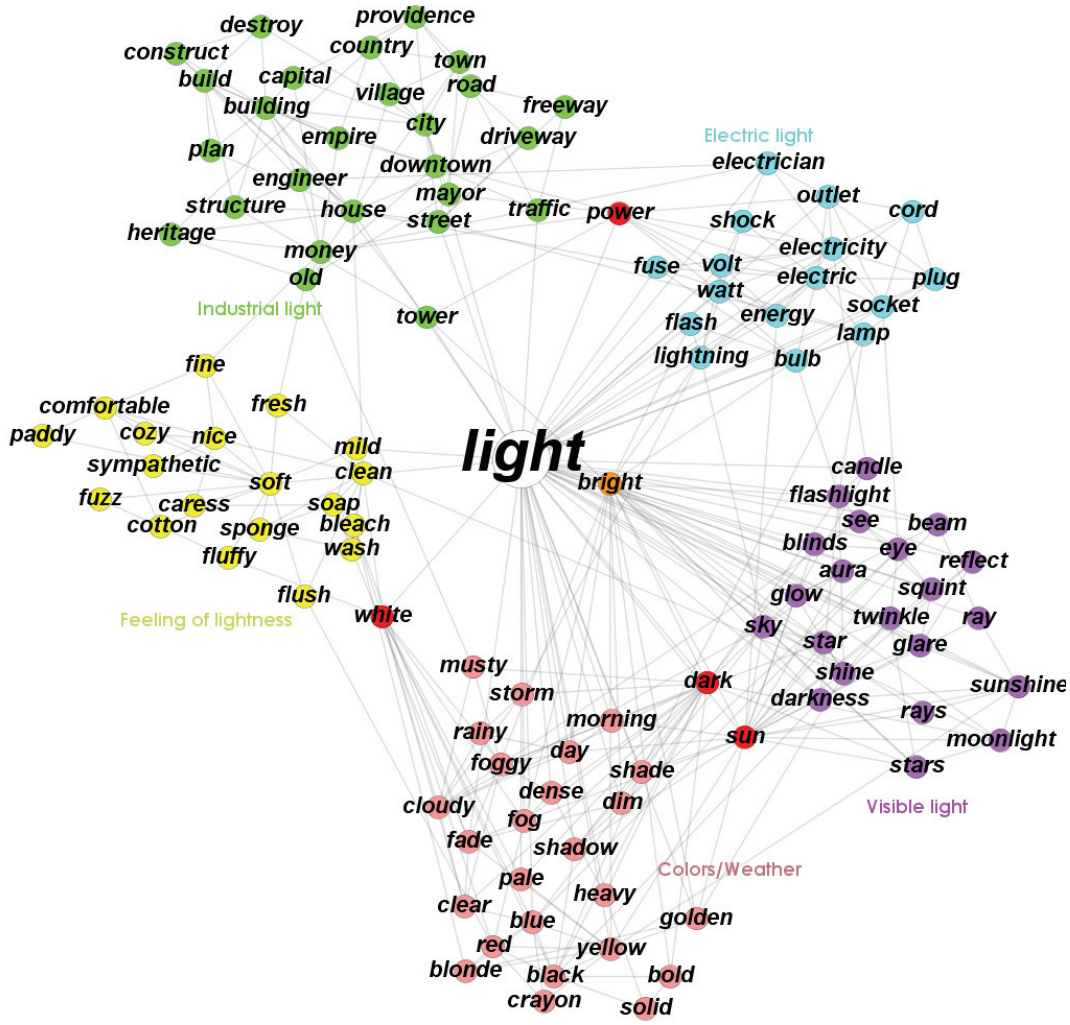


Figure 2.1. An Example Word Association Network. (Source: [26])

2.1.1.2. Collaboration Network

Collaboration is a joint effort of multiple individuals or work groups to accomplish a task or a project. Collaboration, although not a new organizational characteristic, has become a critical factor that determines the success of business [6]. Collaboration network is a network that a group of people or computers or institutions

that combine, help and work with each other and this network is a purpose-built vehicle. Social media analysis is a complex network analysis and the main approach of the researchers is to study and understand relationships and social roles of the people in human communities [6].

In social networks, collaboration networks consist of humans. For instance, in Twitter, people become organized and they can be the trend topic of the day any subject. In the organization process, people have collaboration. An example for a collaboration network is shown in Figure 2.2.

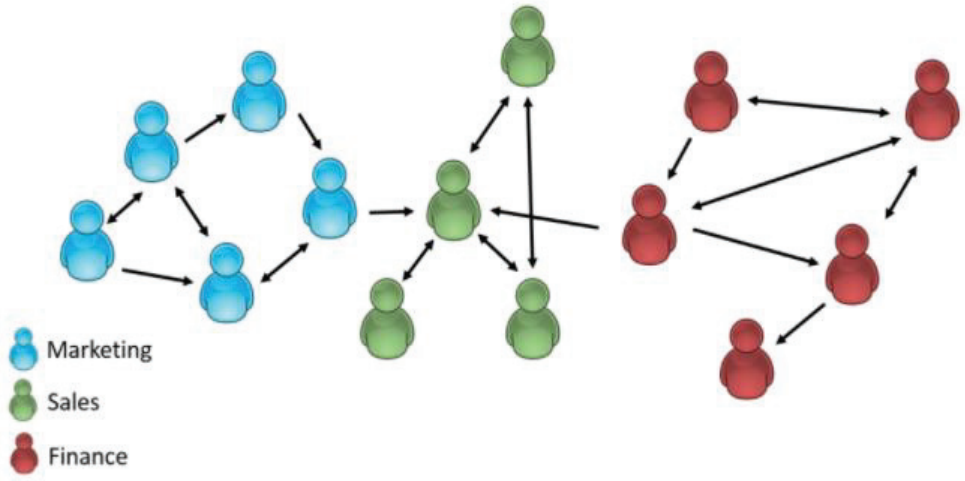


Figure 2.2. An Example Collaboration Network. (Source: [18])

In Figure 2.2., collaboration network has three different communities which are marketing, sales and finance departments. Also, people are the members of the collaboration network. These people in different departments have collaboration to sell more products.

2.2. Community Structure

The property of community structure is that network nodes are joined together in tightly knit groups, between which there are only looser connections [8]. Many systems take the form of networks, sets of nodes or vertices joined together in pairs by links or edges. Examples include social networks such as acquaintance networks and collaboration networks, technological networks such as the Internet, the Worldwide

Web, and power grids, and biological networks such as neural networks, food webs and metabolic networks.

A large body of work in computer science, statistics, applied mathematics, and statistical physics has been devoted to identifying community structure in complex networks [28]. Community structures are interpreted as hierarchical and organizational units in social networks. A schematic structure of communities is given in Figure 2.3.

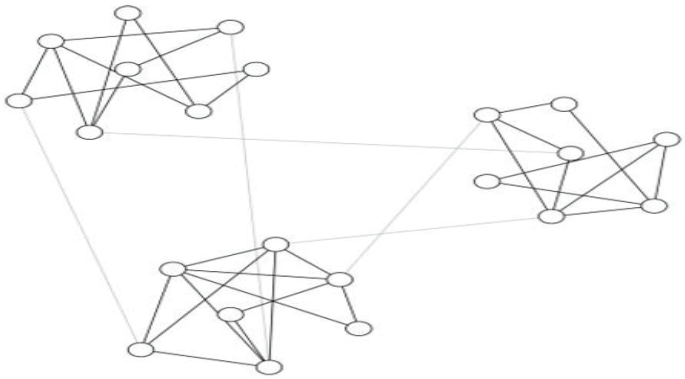


Figure 2.3. A Representation of a Network with Community Structure. (Source: [8])

2.3. Community Detection Algorithms

"*Community structure detection is perhaps best thought of as a data analysis technique used to shed light on the structure of large-scale network data sets, such as social networks, personal interests in social networks, fraud in telecommunication networks, homology in genetic similarity networks and web data, or biochemical networks*" [20]. Community structure methods normally assume that the network of interest divides naturally into subgroups and the experimenter's job is to find those groups [20]. Moreover, community detection methods that are based on density clustering algorithms are only to detect non-overlapping structures [29]. Community structure methods may explicitly admit the possibility that no good division of the network exists, an outcome that is itself considered to be of interest for the light it sheds on the topology of the network [20].

The first community detection algorithms were designed to detect non-overlapping communities in networks. In real and large networks, non-overlapping communities assumption is found to be inadequate. Because, real networks especially,

social networks are so complex networks that nodes have common memberships in different communities.

Many different community detection algorithms have been researched. Some of algorithms are listed below.

- Clustering – Based Methods
- Divisive Algorithms
- Modularity – Based Algorithms

In clustering–based methods, clusters are created by clustering algorithms [10]. A graph can be modelled as an adjacency matrix, and from an adjacency matrix, a Laplacian matrix can be constructed. Spectral clustering can then be applied to the Laplacian Matrix [10].

Divisive algorithms are to detect edges that connect nodes from different communities and remove edges, so that communities are disconnected from the rest of the graph [10]. The most important divisive algorithm is one that is researched by Girvan and Newman [8].

Lastly, modularity is one of the quality functions for communities and it defined in undirected weighted networks as [10]:

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}]\sigma(c_i, c_j)$$

where A_{ij} is the weight of the edge between node i and node j, k_i is the sum of the weights of the edges attached to vertex i, k_j is the sum of the weights of the edges attached to vertex j, c_i is the community that node i belongs to, c_j is the community that node j belongs to, $m = \frac{1}{2}\sum_{ij} A_{ij}$, and $\sigma(c_i, c_j)$ is 1 if node i and node j belong to the same community and 0 otherwise [10].

Modularity is the method for creating a network. In modularity, communities are non-overlapping communities and it is very popular to use complex networks.

2.4. Overlapping Community Detection Algorithm

The technique of detecting overlapping communities in networks is very critical to guide for research of network topology, and the detected nodes by overlapping community detection algorithms belong to multiple communities may play an influential role for network analysis [29].

In social as well as other types of networks, nodes can be a member of multiple communities simultaneously, which leads to overlapping community structures [28]. Such a statement can be demonstrated by the numerous communities each of us belongs to, including those related to our scientific activities or personal life (school, hobby or family) [21]. For instance, people may share the same hobbies in social networks or mechanical engineering and computer engineering students may take the same courses to graduate. It is shown in Figure 2.4.

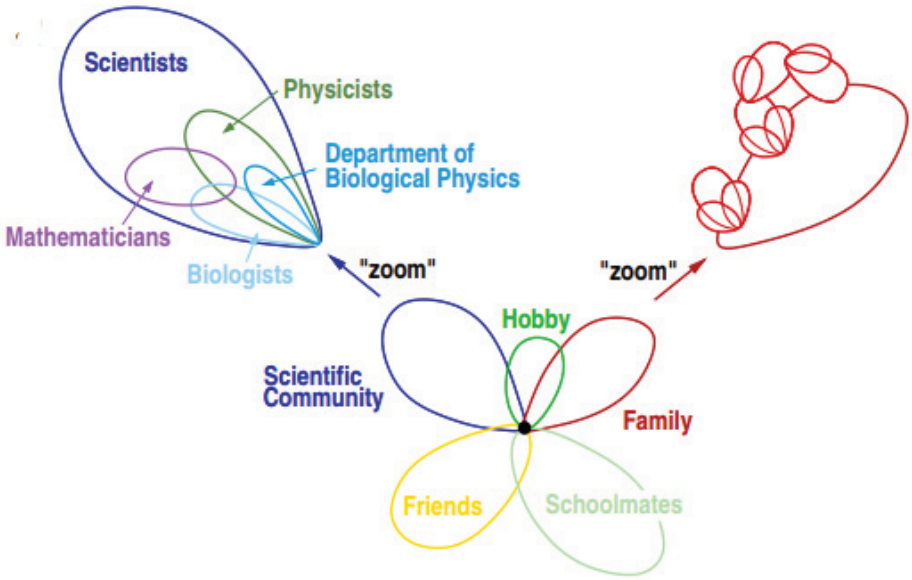


Figure 2.4. Illustration of the Concept of Overlapping Communities. (Source: [21])

BigCLAM algorithm is a [28] state-of-the-art overlapping community detection algorithm and it is designed to find out dense overlaps. Inherently, real social media networks are so complex that they have dense overlaps [28]. In general, community structure may be non-overlapping, sparse overlapping or dense overlapping. An example is given below in Figure 2.5.

In the context of network science, a sparse network is a network with a smaller number of links than the maximum possible number of links within the same network but in dense network, a greater number of links are shared by nodes. BigCLAM is an example of a bipartite affiliation network model [28]. "*Affiliation networks have been extensively studied in sociology as a metaphor of classical social theory concerning the intersection of persons with groups, where it has been recognized that communities arise due to shared group affiliations*" [28].

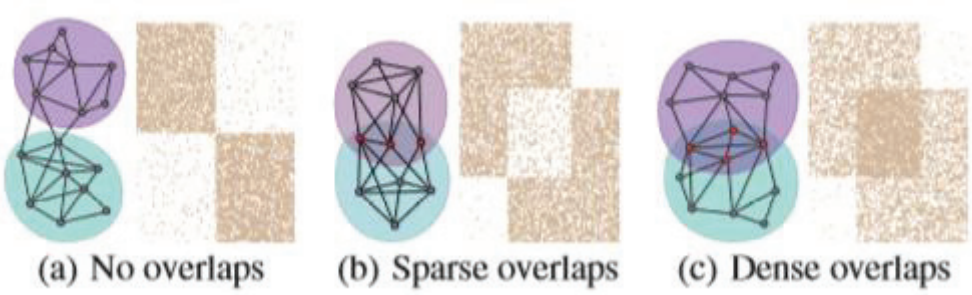
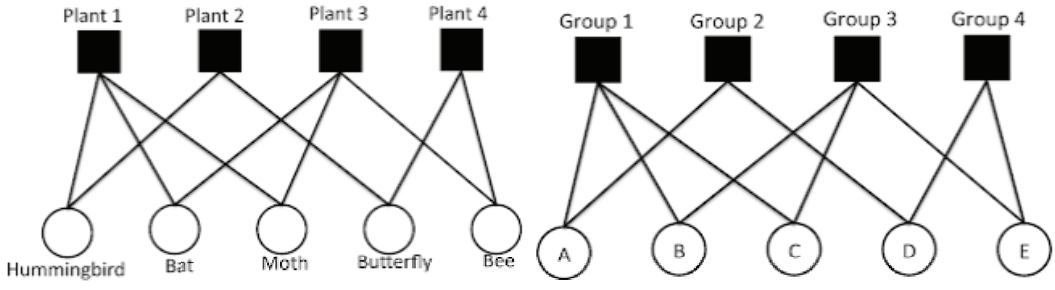


Figure 2.5. Three Views on the Structure of Network Communities. (Source: [28])

In affiliation network models, nodes of the social network are affiliated with communities they belong to and the links of the underlying social network are then derived based on the node community affiliations [28]. Bipartite affiliation network example is given below in Figure 2.6.



A B

Figure 2.6. An Example Bipartite Affiliation Network. (Source: [1])

In Figure 2.6, these two graphs tell the same, but they represent two different scenarios. In A, there are two types of actors and one type of actor can only interact with the other type of actor [1]. Plants have not an edge between them. In B, there are groups and edges represent membership of individuals in groups. These are sometimes

called affiliation networks. Individuals being one set of nodes, and groups being another set of nodes. Individuals can only be connected to groups, and vice versa [1].

BigCLAM formulates community detection as a variant of non-negative matrix factorization [28]. BigCLAM algorithm has two important improvements. Firstly, a lot of non-negative matrix factorization algorithms pay relatively less attention to interpret the latent factors [28]. The primary goal there is to estimate the missing entries of the matrix [28]. The second improvement is to use Gaussian distribution or logistic link function optimized the model likelihood of explaining the links of the observed networks [28]. "*In practice, computing the gradient in near-constant time makes the BigCLAM algorithm about 1000 times faster*" [28]. Also, in terms of scalability, BigCLAM algorithm has some advantages. In general, most overlapping community detection methods scale to networks with at most thousands of nodes [28]. For instance, mobile phone network is the largest network and it has 800000 nodes and 2.8 million edges [28]. "*BigCLAM can process networks with tens of millions of edges while also obtaining state of the art quality of detected communities*" [28].

In BigCLAM algorithm, networks can be modelled with different types. For instance, our network may be a social media network and just non-overlapping network communities are not enough to reach the correct results. BigCLAM supports non-overlapping, overlapping or nested community models. These models are figured out in Figure 2.7.

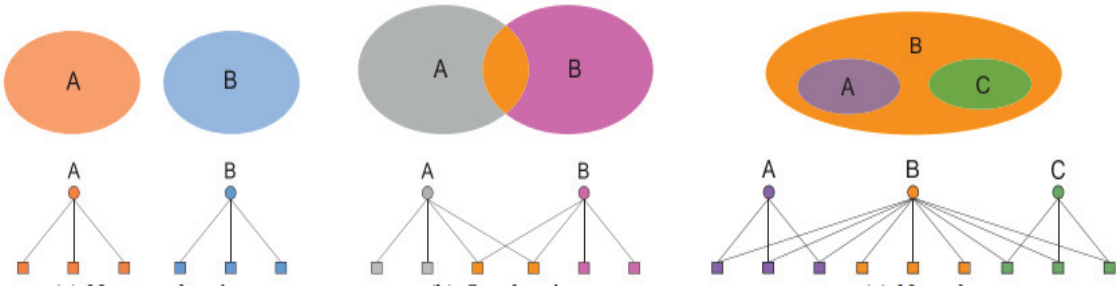


Figure 2.7. Three Models of Network Communities.

Clique percolation algorithm is one of the basic algorithms in detection of overlapping communities.

In clique percolation, "*the basic observation on which community definition relies is that a typical community consists of several complete (fully connected) subgraphs that tend to share many of their nodes*" [21]. Thus, defining a community, or

more precisely, a k-clique-community as a union of all k-cliques that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing $k-1$ nodes) [21]. Community members can be reached to subsets of nodes [21]. The other parts of the network are not reachable from a k-clique, but they have a potential to contain further k-clique communities [21]. As a result, a node can belong to more than one communities. That is the logic of the overlapping community structures.

2.5. Natural Language Preprocessing

"*Information retrieval (IR) is finding material (usually documents) of unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)*" [19]. The success of an information retrieval system is assigned based on how it can minimize (compress) the information, the amount of related information extracted for user needs and how fast it retrieves information before presenting to user. Ancestor of information retrieval systems can be accepted as hard-copy catalogues and central repositories [14].

Information storages in many different environments have increased continuously and these storages demand to be analyzed. NLP functions are used to correct the information and to make it easy to analyze. In this section, some of the NLP functions will be explained.

2.5.1. Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation [26]. For example, a sentence is split into words or phrases. Also, tokenization process removes punctuations, capitalizations and all the changes / modifications from sentences.

2.5.2. Stop Words Removal

Stop words are common words which generally occur in all documents, therefore in most cases they do not imply specific semantic meaning in document content [19]. These words can be given as examples for stop words: “of”, “at”, “a”, “or”, “and”.

Stop words are used in nearly all documents very frequently and size of stop words is very big in many corpora that’s why stop words removal has some advantages. The most important advantage is to reduce the size of index of documents and it leads to decrease the computation time because of reduction in word count. Also, it is important to determine word similarity correctly. Word similarity calculation is done with word sighting in specific window size.

2.5.3. Stemming

Stemming is a grammatical modification of a word by using some defined rules, generally by chopping off the end of the word [19]. Stemming is used to identify the same word out of two different words that have similar root or origin. For example, in Turkish, “eğitim” and “eğitimi” words have the same root that is “eğitim”. These two words cannot be evaluated as different words. The aim of the stemming operation is to process the word and reach to the ground forms.

2.6. Statistical Semantics

Semantics is determining the meaning of texts, books or sentences and every natural language has a semantic structure. Statistical semantics is the study of estimation of the meanings of words by looking at patterns of words using statistical methods in huge collections of texts. Nowadays, extraction of usable information is very important from social network. Social network has many information about many people and if used words can be analyzed semantically, this information can be used in many

different areas such as marketing, market analysis, and advertising. Statistical semantic processes can be used for extracting useful information from social media.

2.6.1. Co-occurrence Matrix

"*Within social sciences, word co-occurrence analysis is widely used in various forms of research concerning the domains of content analysis, text mining, construction of thesauri and ontologies*" [15]. In general, the aim of co-occurrence matrix is to find similarities in meaning between word pairs and/or similarities in meaning among/within word patterns, also, discovering the latent structures of mental and social representations [15].

In word similarity, a co-occurrence matrix consists of words and it can be a rectangular or square matrix. The purpose of this matrix is to present the number of times each row word appears in the same context as each column word. The co-occurrence matrix example is shown in Figure 2.8 and the sentence input with respect to which the matrix is created is as follows:

"*Did you read this book? You can read the latest book of Harris to improve your work.*"

	read	you	book	did	this	can	the	latest	of	Harris	to	improve	your	work
read	0	0	1	0	1	0	1	1	0	0	0	0	0	0
you	2	0	0	0	1	1	0	0	0	0	0	0	0	0
book	0	1	0	0	0	1	0	0	0	0	0	0	0	0
did	1	1	0	0	0	0	0	0	0	0	0	0	0	0
this	0	1	1	0	0	0	0	0	0	0	0	0	0	0
can	1	0	0	0	0	0	1	0	0	0	0	0	0	0
the	0	0	1	0	0	0	0	1	0	0	0	0	0	0
latest	0	0	1	0	0	0	0	0	1	0	0	0	0	0
of	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harris	0	0	0	0	0	0	0	0	0	0	1	1	0	0
to	0	0	0	0	0	0	0	0	0	0	0	1	1	0
improve	0	0	0	0	0	0	0	0	0	0	0	0	0	1
your	0	0	0	0	0	0	0	0	0	0	0	0	0	0
work	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2.8. Co-Occurrence Matrix Example (window size = 2)

The matrix in Figure 2.8 is a square co-occurrence matrix. Co-occurrence matrix creation process starts by splitting the sentences into words. After that, all words in the corpora are scanned and a co-occurrence matrix is created with these words which occur in a specific window-size together. Two (or more) words that tend to occur in similar linguistic contexts (i.e. to have similar co-occurrence patterns), tend to be positioned

closer together in semantic space [15]. The reason why these kinds of outputs are so popular depends on the fact that multidimensional analysis allows us to represent the entire structure of data matrices and allows us to discover new information patterns by highlighting similarities and differences between objects [15].

2.6.2. Normalization

This part presents the normalization calculations. Normalizations are used on co-occurrence matrix results. There are two different normalizations which are Positive Pointwise Mutual Information (PPMI) and Alternate Pointwise Mutual Information.

2.6.2.1. Positive Pointwise Mutual Information

Pointwise Mutual Information [27] is a measure of how much the actual probability of a co-occurrence of events $p(x, y)$ differs from what we would expect it to be on the basis of the probabilities of the individual events $p(x), p(y)$ and the assumption of independence. Positive Pointwise Mutual Information is extracted from pointwise mutual information, its results are non-negative. The formula of Pointwise Mutual Information is given below.

$$pmi(x, y) = log \frac{P(x, y)}{P(x)P(y)} = log \frac{P(x|y)}{P(x)}$$

where;

$P(x, y)$: joint probability of words x and y in the same window size

$P(x)$: probability of the occurrence of x in the corpora

$P(y)$: probability of the occurrence of y in the corpora

In Positive Pointwise Mutual Information calculation, if the result of the formula is negative or undefined the result is assumed to be zero.

2.6.2.2. Alternate Pointwise Mutual Information

Alternate Pointwise Mutual Information is another calculation method to use as the measure of association of words in matrix. The formula of Alternate Pointwise Mutual Information is given below.

$$pmi(w,c) = log \frac{(w,c) * D}{w * c}$$

- Let's assume that index $w = 2$ and index $c = 3$
- (w,c) : A value in co-occurrence matrix cell indexed into $(2,3)$.
- Value w : In the second row, addition of all values except $(2,3)$ cell value.
- Value c : In the third row, addition of all values except $(3,2)$ cell value.
- D : The number of words in corpora.

There are two different pointwise mutual information calculations, and both are trying to reach the same result. In the first formula, $P(x,y)$ returns the joint probability of x and y on a specific window size. After the simplification of the formula the numerator turns into sighting of x and y on a specific windows size multiplied with total word size in corpora. In the second formula, $(w,c) * D$ returns the same. In denominator part, first formula gets the total occurrence count of x and y. In the second formula, sighting of x and y count is subtracted from the total occurrence count of x and y separately. That is the main difference of these two pointwise mutual information calculations.

2.6.3. Similarity Measurement

Similarity measurements are used for calculating the similarity of some objects or groups. Cosine similarity and Jaccard similarity measures will be described as they were appropriate to be used in the context of this thesis.

2.6.3.1. Cosine Similarity

Cosine similarity function, which is the measure of similarity between two vectors derived from the cosine of the angle between them is used for finding similarities of words in text documents. In space, words can be transformed into vectors and the cosine value between these two vectors gives a measure of similarity. Such functions have largely been used in the web space to identify similarity of text documents and web pages. It has been one of the most preferred techniques in information retrieval, clustering and is even applied to pattern recognition and medical diagnosis [23].

Formally, before all, the dot product definition is given for two vectors $\vec{a} = (a_1, a_2, a_3, \dots)$ and $\vec{b} = (b_1, b_2, b_3, \dots)$ where a_n and b_n are the components of the vector and n is the dimension of the vectors. The dot product formula is given below:

$$\vec{a} \,.\, \vec{b} = \sum_{i=1}^{n} a_i b_i$$

The cosine similarity shows the angle between two words. Cosine similarity formula is given below.

$$cos\theta = \frac{\vec{a} \,.\, \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|}$$

Note that since the cosine similarity function measures the angle between two vectors, it is a symmetric similarity function [3].

In a co-occurrence matrix, words settle into rows and columns and they have values in their cells. Cosine similarity is calculated for all rows in a co-occurrence matrix one by one. Values of every cell create the cosine similarity function inputs and determine the cosine similarity value of two rows.

2.6.3.2. Jaccard Similarity

The Jaccard similarity [12] (also called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It is a measure of similarity of two sets of data with a range from 0 to 1. In Jaccard similarity, if two datasets are similar then the Jaccard similarity of two datasets will be approached to 1. The Jaccard similarity formula is given below.

$$J\,(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

where;

- $|X \cap Y|$ is the absolute value of intersection of X and Y.
- $|X \cup Y|$ is the absolute value of union of X and Y.

For example, there is big text corpora and let's x and y become users.

- User x writes in 85 different topics.
- User y writes in 60 different topics.
- User x and user y write in 25 similar topics.
- $J\,(X,Y) = \frac{25}{(85+60)-25} = 0{,}2083$

2.6.4. Evaluation Criteria

Evaluation criteria are used for comparing cluster quality and similarity values. Networks have communities and the similarity of communities determine the similarity of networks. Normalized mutual information, rand index and F-score calculation can be used to compute the similarity of networks.

2.6.4.1. Normalized Mutual Information

Normalized Mutual Information (NMI) is a good metric for determining the quality of clustering [25]. In NMI, result can be between 0 and 1 and if the result is near to 1 then the similarity of clusters is high and if the result is near to 0 then the similarity of clusters is low. NMI is normalized and we can compare clustering with different number of clusters. The NMI formula is given below:

$$NMI(\alpha, \beta) = \frac{I(\alpha; \beta)}{[H(\alpha) + H(\beta)]/2}$$

where;

- α and β are parameters of the formula for Normalized Mutual Information.
- I is mutual information and it measures the amount of information by which our knowledge about the classes increases when we are told what the clusters are. Minimum value of I is 0.
- H is entropy and defined as follows:

$$H(P) = -k_B \sum_i p_i \log p_i$$

2.6.4.2. Rand Index

Rand Index (RI) is an alternative way to measure of clustering quality [11]. In prediction models we have 4 parameters which are given below in Table 2.1:

Table 2.1 Prediction Model Attributes

	Predicted Class		
Actual Class		Class = YES	Class = NO
	Class = YES	True Positive	False Negative
	Class = NO	False Positive	True Negative

In this table, 4 attributes are shown. Descriptions are given below:

- True Positive: True positive means that the result of actual class is yes, and the prediction algorithm predicts also yes. This is a correct prediction.
- True Negative: True negative means that the result of actual class is no, and the prediction algorithm predicts also no. This is a correct prediction.
- False Positive: False positive means that the result of actual class is no, and the prediction algorithm predicts yes. This is a false prediction.
- False Negative: False negative means that the result of actual class is yes, and the prediction algorithm predicts no. This is a false prediction.

In our study, a true positive (TP) decision assigns two similar words to the same cluster, a true negative (TN) decision assigns two dissimilar words to different clusters. There may be two types of errors. One of them is a false positive (FP) decision which assigns two dissimilar words to the same cluster. The other is a false negative (FN) decision which assigns two similar words to different clusters.

The Rand Index [11] measures the percentage of decisions that are correct. The Rand Index formula is given below:

$$Rand\ Index\ (RI) = \frac{TP + TN}{TP + FP + FN + TN}$$

For example; we are computed some values. True positive is 40, false negative 24, false positive 20 and true negative 68. Then the Rand Index value is calculated like this:

$$RI = \frac{40 + 68}{40 + 20 + 24 + 68} = 0.7105$$

2.6.4.3. F1-Score Calculation

Data science is a very important research area nowadays. Today, we have large amounts of data to analyze and extract meaningful result. Scientists put a model to analyze the data and the most important aim of the model is to have high rate predictability. F1-score calculation can be used to measure predictability.

F1-score [2] calculation needs precision and recall values. Descriptions of precision, recall, and F1-score are given below:

- Precision: Precision is the ratio of correctly predicted positive values to the total predicted positive values.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Recall is the ratio of correctly predicted positive values to the all values in actual class which have yes values.

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score: F1-score is the harmonic average of precision and recall values. If F1-score is near to 1 then prediction is good for this scenario.

$$F1 - Score = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

The Rand Index gives equal weight to false positive and false negative values. Separation of similar words may be worse than putting pairs of dissimilar words in the same cluster.

F-score formula can be modified if we penalize false negatives more strongly than false positives by selecting a value more than 1.

$$F_\beta = \frac{(\beta^2 + 1) * (Precision * Recall)}{\beta^2 * Precision + Recall}$$

CHAPTER 3

RELATED WORK

This chapter presents the related work about this research. There are various works about social media and network analysis. We focus particularly on community detection on online social networks.

Node connections are showed having relations such as web page relations, social relations, and biological relations, which are complex networks [7]. There are many community detection algorithms and some of them can detect overlapping communities, some of them detect non-overlapping communities and some of them detect both. Non-overlapping community detection using local community approaches is studied [7]. In this study, the approach uses similarity measurements to determine community of nodes. Average neighborhood ratio is computed. Secondly, similarity measure is computed for pair of nodes suspected to be in the same community [7]. Nodes which are qualified are merged into a local community. After that, node pairs which are placed in local communities are merged into the global communities. As a result of the work, the proposed algorithm compares with existing algorithms and it shows that accuracy and prediction rate of communities is high [7].

The definition of community is given as follows: "*A community is typically thought of as a group of nodes with more connections amongst its members than between its members and the remainder of the network*" [28]. There are many difficulties to detect communities in large and complex networks. Create a new algorithm called BigCLAM, which is an overlapping community detection algorithm and it can be run in large networks which have millions of nodes and edges [28].

In the impact of collaboration and knowledge networks on citations [9] research is on collaboration and knowledge networks together. The definition of a collaboration network is given as "*a typical social-based network in research*". In these type of networks, the attributes of authors play important roles for instance in this research, citation count affected authors' centrality positively. The definition of a knowledge network is: "*A knowledge network is comprised of combinations between components or elements of scientific or technological knowledge*" [9]. Scientific papers on specific

subjects are studied more than one keywords are used to relate them. These relevant keywords create a co-keyword network to represent the knowledge network. These keywords are connected through their co-occurrence with each other. In this research, the structural attributes of knowledge elements in the network will influence the nodes' combination opportunities and efficiency [9]. For example, a node which is a central element of the network is more likely to be searched and easily combined with other nodes because it has more contents of element couplings. Also, knowledge network elements may affect the citation network. In this study, the knowledge network is created and the influence of this network on paper citations is investigated.

In information cartography, the problem is defined as extracting structured knowledge automatically from very large datasets [24]. In this study, metro maps are created automatically from large datasets and the user can understand the flow of events easily. When the event is created, many of sub events may be triggered of this. In big picture, the user can see many metro lines which are sub events of the main event and these metro lines may be intersected or overlapped [24]. The metro stops create a cluster of articles. In this research, metro maps are described as: "*Each line follows a coherent narrative thread, and different lines focus on different aspects of the story. Intersections across lines reveal the ways different storylines interact.*" [24]. Coherence is recognized as the first important thing because each metro line tells a coherent story to the user. In chain of clusters, each cluster is created from the set of documents. In information cartography research, the concept of coherence is represented as such: "*In order to define coherence, a natural first step is to measure similarity between each two consecutive articles along the chain*" [24]. After that, problem is transformed into a linear programming problem where the aim is to choose a small set of words and giving score the chain is based solely on these words. The score of the chain equals to the weakest link in the chain because the strength of the chain is measured with the strength of the weakest link in the chain [24]. Coverage is the second important characteristic because the event is shown fully by a balance of coverage and coherence for extracting a meaningful map. Firstly, the set of elements is defined, and these elements must be covered for extracting a good map. Elements may be words. After that, coverage function is run, and it calculates how well each document covers each element and good documents are picked up [24]. In this research, elements are weighted with a natural mechanism for personalization. After that, set of lines are connected and this represents connectivity for this study [24].

Social networks try to be represented as a network graph with overlapping community detection algorithms. These network communities have densely connected nodes and overlapping community structure gives an opportunity to nodes which are members of more than one community [4]. Nodes and edges are described as: "*In social networks, nodes represent users/actors/items whereas edge represents the link/flow of interaction/relationships among the users*" [4]. In social networks, each node represents entity and analyzing of network links gives the behavior rules or patterns about different activities and detecting communities involve clustering of similar users in social network graph. In this research, community description is given as: "*Community is a module containing the set of nodes with major activities/interaction/similarity among them*" [4]. First two disjoint communities are identified with Newman's modularity and the modularity value is found near to 1. Other modularity algorithms are compared with each other. After that, the same data creates two communities which are overlapped, and one node participates in both two communities. In this research, the strength of the participation is measured with "*belonging degree*" and they say "*Its value differs according to each researcher's view and may depend on the domain/application chosen*" [4].

CHAPTER 4

EXPERIMENTAL WORK

This chapter presents the experimental process in steps. First, data collection process from the real, Turkish social media is described. In social media, people do not pay attention to what they write that's why social media data have many grammatical errors and nonsensical words. Consequently, this data is preprocessed by the natural language processing techniques of tokenization, stop word removing, and stemming, to give consistent results.

Third, processed social media data is transformed into a co-occurrence matrix to be ready for similarity measurement and word association network and collaboration network data are created by similarity calculations. Fourth, the networks are processed by the BigCLAM overlapping community detection algorithm. Then, extracted communities are analyzed and visualizations are prepared. The experimental workflow is given below in Figure 4.1.

In workflow, the first process is data collection. The data is crawled via crawler which is coded in Java. After that, the data were separated by meaningful parts and stored in relational database. This part was database management process. After the data collection, Zemberek was integrated into the project and NLP operation processes were started. In this part, three operations were performed. These were tokenization, stop word removal and stemming. After the NLP operations, co-occurrence matrix was created. While creating operations, some mathematical similarity calculations were used. These were pointwise mutual information and alternate pointwise mutual information. Then cosine similarity was calculated for co-occurrence matrix and word association network was created. After that, collaboration network was created with Jaccard similarity calculations. Lastly, these networks were analyzed in graphs and overlapping, and non-overlapping networks were compared with each other.

Ekşisözlük is a very popular Turkish social media platform and people share their ideas about daily life, trends, politics, and some specific subjects.

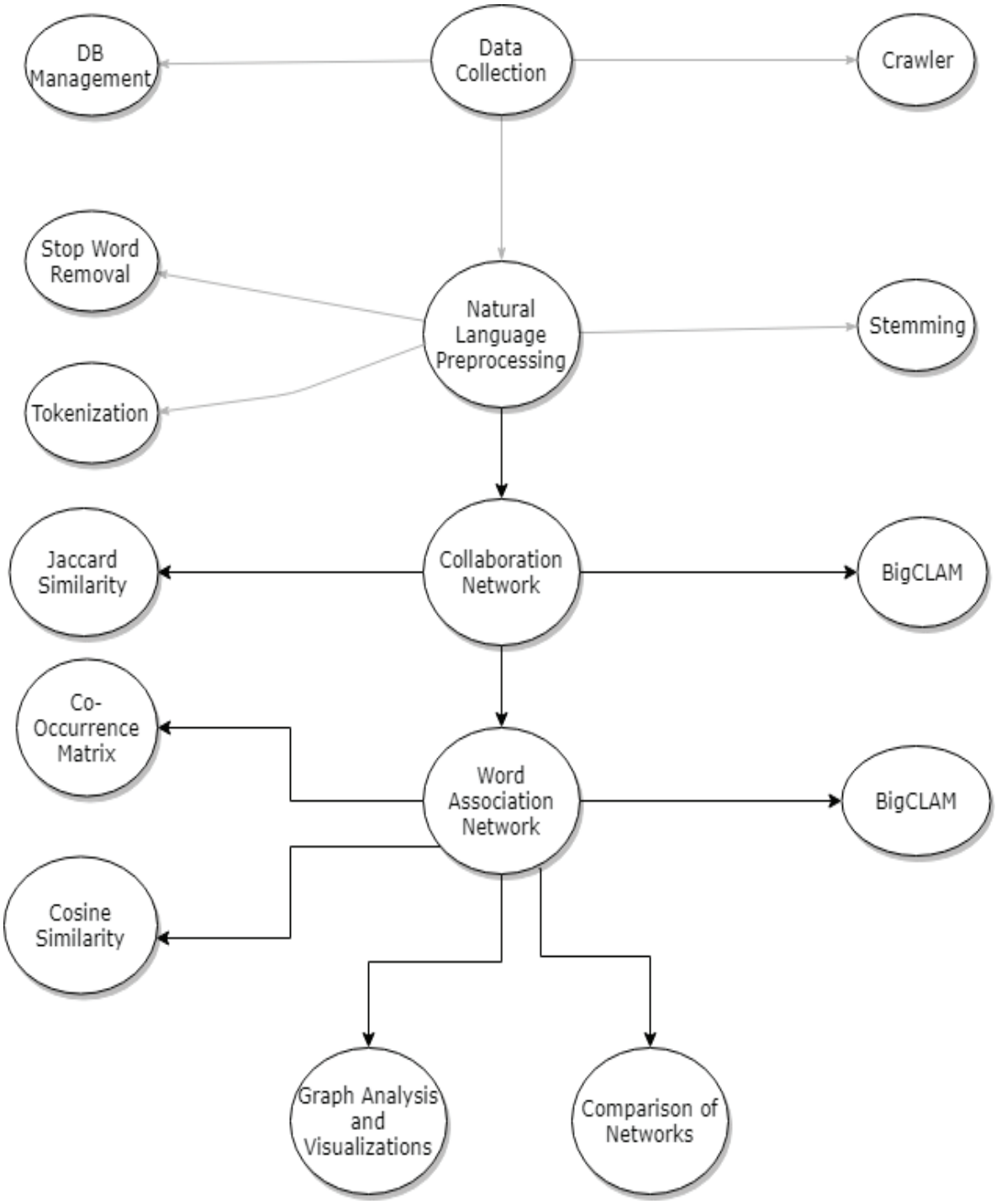


Figure 4.1. The Experimental Workflow.

Ekşisözlük works with a membership system. Before all, people must be members of Ekşisözlük after that they can make some comments about titles. Ekşisözlük has many structures but basically a few critical and important structures for this research are title, entry, and user. Description of these structures are given below.

- Title: Title structure represents the subject and gives general information about comments of users. Title structure provides a grouping of entries. Users create titles and other users write their comments about the title.
- Entry: Entry structure is a comment of a user about a title. One entry is written into only one title and it is written by only one user. All titles must be written lowercase. Entry is to be evaluated like a tweet for Twitter in Ekşisözlük.
- User: Users write and read entries. Some users are visitors and they only can read the entries. Some users are registered, and they are authors. All users can read the entry. Some users are registered, and they are rookies. They can write about title, but their entries are not kept out of sight.

4.1. Data Collection

Some social media sites have private accounts to publish shared things and some social media sites have public accounts and public sharing of things. These are security configurations and are configurable for users. For instance, in Twitter, users can lock their accounts and only followers can see the tweets. In Ekşisözlük, all entries are visible and everyone can see what is shared. Also, Ekşisözlük is a social environment, which has content in Turkish.

Firstly, in Ekşisözlük, every title has unique link that's why title links are very important to reach the entry information. Ekşisözlük has sitemap.xml file and everyone can reach this xml file with link as given below and the sample file is shown in Figure 4.2.

https://eksisozluk.com/sitemap.xml

This file has $< url >$ tags and these tags have $< loc >$ tags. These $< loc >$ tags contain URL of entries. The $< lastmod >$ tags are not important for collecting title links. After the collection process, URLs are separated into meaningful parts with a rule-based approach. Title URL of Ekşisözlük separation rules are given below:

- Domain => eksisozluk.com
- Title information => /yalniz-yasamak—156903
- Pagination =>?$p = \{1, 2, 3, 4, 5, \dots, n\}$

```
<url>
<loc>https://eksisozluk.com/istatistik</loc>
</url>
<url>
<loc>
https://eksisozluk.com/ziraat-bankasinin-kuluplerin-borclarini-odemesi--5897791
</loc>
<lastmod>2019-01-06</lastmod>
</url>
<url>
<loc>https://eksisozluk.com/eksi-itiraf--1037199</loc>
<lastmod>2019-01-06</lastmod>
</url>
<url>
<loc>
https://eksisozluk.com/evrimcilerin-dustukleri-en-buyuk-hata--5898908
</loc>
<lastmod>2019-01-06</lastmod>
</url>
<url>
<loc>
https://eksisozluk.com/konyadaki-ilginc-mezar-tasi--5898898
</loc>
<lastmod>2019-01-06</lastmod>
</url>
<url>
<loc>https://eksisozluk.com/lojistik-yonetimi--1017617</loc>
<lastmod>2019-01-06</lastmod>
</url>
```

Figure 4.2. The Sample File for sitemap.xml.

Title information splits into words and there is a hyphen for each word. At the end of the title information, the id of the title specifies the link and there are three hyphens between id and the last word. After the title information, the link continues with pagination part with question mark and numbers. If pagination number increases, then more social media data can be reachable for the given title.

Secondly, after the analysis of Ekşisözlük structure, a crawler is developed. This crawler gets information in HTML format and separates every information rule-based. After that, the data from social media must be processed correctly. The data are processed with rules and every smallest information source has attributes which are given below:

- Title information of the entry
- Entry content
- Entry id
- Entry link
- User information of the entry
- Like count of the entry

- Unlike count of the entry

The title information of the entry gives the general information about the entry content and it is information for grouping entries. Entry contents are the origin of this research because this information will be analyzed with natural language processing operations and it will be processed with similarity operations to create word association networks. User information of the entry gives the user name of the user who writes the entry. Also, user information can be used for grouping entries for creating collaboration networks. Like and unlike count of the entry come with the information coming from the website and it is a useful information to give an idea about the popularity of entry of the user. The entry link and user names are unique information in the system.

Our research data are social media data and have a lot of information and hard to analyze. That's why, to process and analyze them easily, all data are stored in a relational database system. The ER diagram for database design is given below in Figure 4.3:

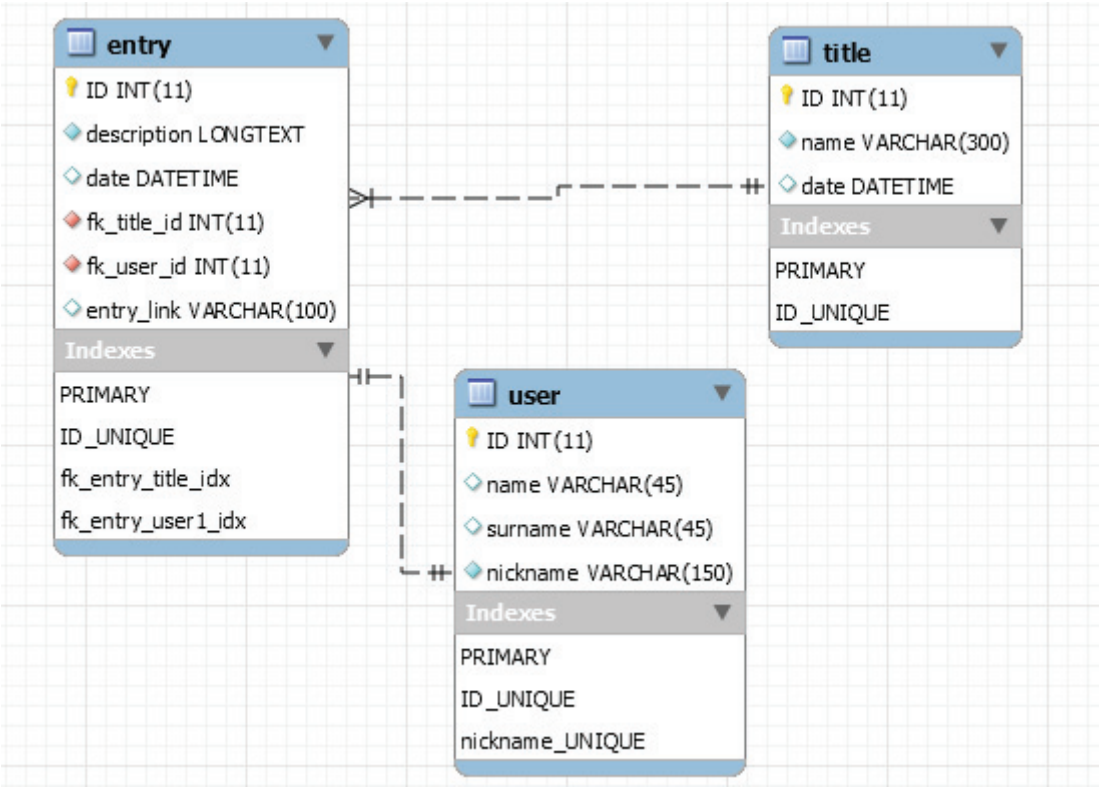


Figure 4.3. The ER Diagram of Relational Database

By the help of a relational database, many different queries and many different analyses are run. Approximately we analyzed results which are given below.

- 8000 titles
- 2.000.000 entries
- 82.000 information of users are crawled from Ekşisözlük and analyzed

4.2. Natural Language Processing

Natural Language Preprocessing operations are applied for correcting and simplifying words in documents, articles, social media etc. NLP operations are the same for every language, but application of rules varies by language. There are some libraries to do NLP in Turkish. Zemberek library is a Natural Language Processing library for Turkish language.[1]

Zemberek project is integrated to this research as a module and NLP operations are performed with this library. Before all, all entries are grouped by titles and users separately. After that, these data are extracted as an input file for NLP operations. The workflow of NLP operations is shown in Figure 4.4 below.

Firstly, there are many entries in the database and they must be grouped to be analyzed with NLP operations easily that's why entries are grouped by titles, after that the same entries are grouped by users. These grouped entries are exported as .txt files from database and they will be input for NLP functions. In workflow, functions are shown as circles and they are lined up according to working order.

In the first step, tokenization operations are applied to all entries. Entries are separated into sentences after that sentences are separated into words. Punctuations and capitalizations are removed.

Secondly, stop words are used in many sentences and they do not have any semantic meaning while using in sentences sequestered that's why stop words were removed from corpora and they were not calculated similarity values by similarity calculations. In many applications, stop words are read from an external document. Different research groups published some static documents for the list of stop words. System reads this external document and scans all words from corpora and if it detects

[1] Available on: https://github.com/ahmetaa/zemberek-nlp

the stop word then that word is removed from the document. Thirdly, Zemberek library has a normalization function for Turkish words. Especially, in social media many words may be written wrongly. For instance, social media user may write “Ankraa”, but user wants to state “Ankara” actually. For increasing the correction results of similarity calculations, all words are analyzed one by one in the corpora and if the normalization function decides that the word is wrong then the first suggestion of the normalization function is replaced with the wrong word. If the list of suggestions is returned null, then we say the word has no misspelling.

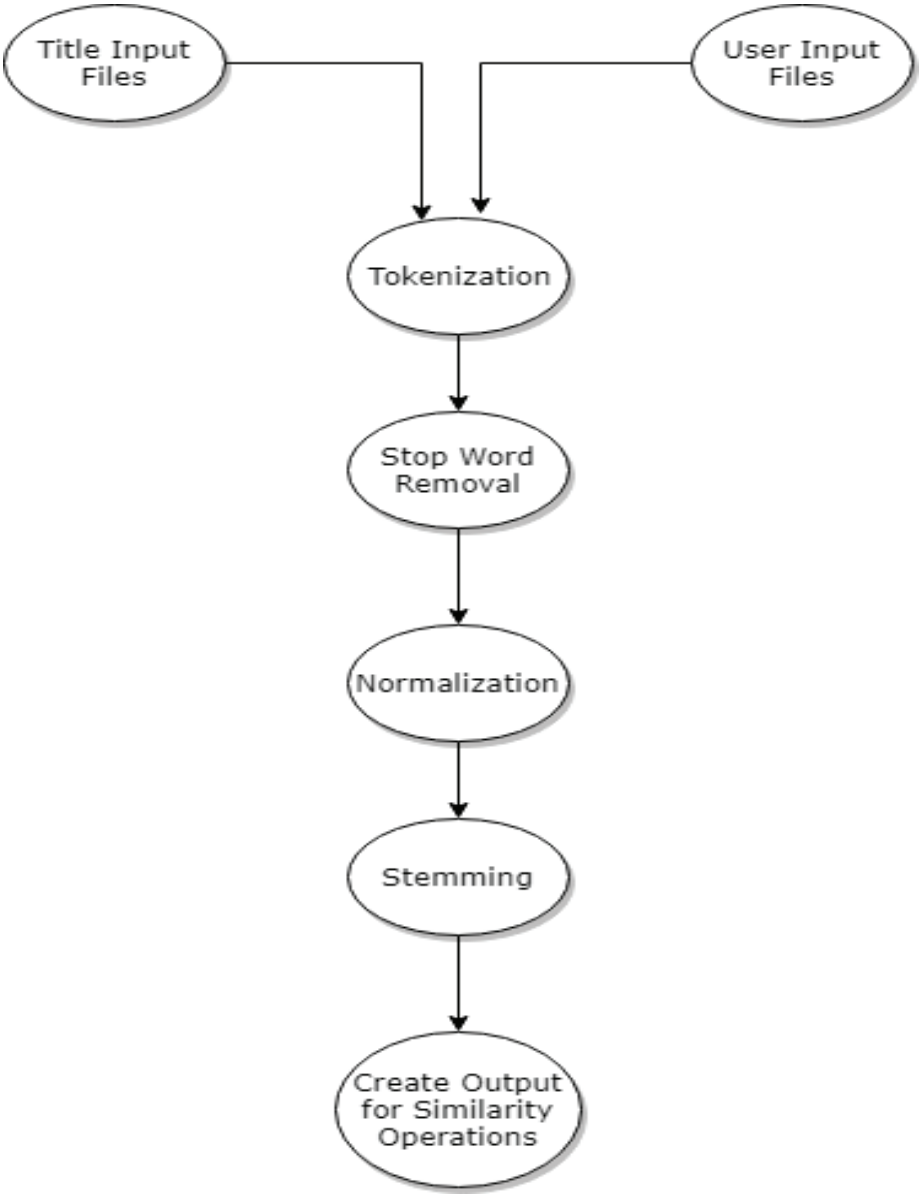


Figure 4.4. The NLP Operations Workflow

In speaking and writing, words have suffixations. For example, in Turkish “arabaya” and “arabayı” words are derived from the root “araba” and these two words

cannot be counted as two different words that's why after the normalization operations, stemming is done as a next step. All words are analyzed and changed with their roots all along the corpora. This operation is the fourth operation of natural language preprocessing operations.

Lastly, after performing all NLP operations, the output files are created for all titles and users. These files are given to a co-occurrence matrix and network creation functions as an input.

4.3. Collaboration Network, Co-Occurrence Matrix and Word Association Creation Processes

Networks are used to analyze complex structures easily. Different phenomena can be represented by a network structure. For instance, in social media, nodes can represent users, videos, pictures, words etc. Another example can be given from a sales network, where nodes may represent customers or products. Links of nodes give an information about relationships of nodes.

In this research, word association network and collaboration network are created from the same data. A word association network represents the similarity relationships of words whereas a collaboration network represents relationships among social media users.

4.3.1. Collaboration Network Creation Process

In Ekşisözlük, users write an entry and this entry belongs to a title. Some users write an entry to the same title more frequently. In Ekşisözlük, titles represent a subject about sport, politics, or daily life. In the light of all information, a social media network can be created whose nodes consist of real social media users. These nodes have links with each other and these links represent the relationships of social media users. The network which represents relationships among social media users is termed as a collaboration network.

In this research, in defining relationships among social media users, writing an entry to the same title is referred. First, information on how many different entries each

user entered in different topics was calculated. This result is retrieved from the database with a suitable query. First 1000 users are retrieved from the database and they are listed in Table 4.1 which is named as top users contributing to different entries. The small example is given below in Table 4.1:

Table 4.1. A Sample from Top User-Title Table

User name	Title Count
User1	593
User2	586
User3	543
User4	538
User5	510
User6	506
User7	492
User8	466
User9	462
User10	453

In this table, the first column shows the embowered version of usernames and the second column shows the different title count that the user contributed.

Secondly, top users who write most entries to different titles resemble a certain extent. After depicting the number of different entries each user entered in different topics, two users are associated by how many titles in common they write into. User-user-title similarity sample is shown in Table 4.2 and examples are given below.

After that, the similarity of two different users is calculated by Jaccard similarity. Jaccard similarity is a measure of similarity of two sets of data with a range from 0 to 1. Sets consist of user-title relations. A sample of Jaccard similarity values is given below n Table 4.3.

Jaccard similarity calculation is used for calculating user similarities. If Jaccard similarity value is close to 1, these users are similar. Based on the Jaccard similarity values for users, a collaboration network is created. As the resultant collaboration network is huge and hard to be analyzed with tools, output is pruned with quantile calculation.

Quantile calculation shows information about the output distribution. Jaccard similarity values are the input of quantile calculation. After the calculation, distribution is given below in Table 4.4.

Table 4.2. A Sample from User-User-Title Similarity Table.

User name 1	User name 2	Similar Title Count
User1	User2	110
User1	User3	197
User1	User5	68
User4	User1	108
User8	User1	98
User4	User5	47
User7	User9	84
User10	User3	155
User6	User4	29
User8	User7	102

Example 1:

- User1 writes in 593 different titles from user-title table.
- User2 writes in 586 different titles from user-title table.
- 110 titles are common for User1 and User2.

Example 2:

- User4 writes in 538 different titles from user-title table.
- User1 writes in 593 different titles from user-title table.
- 108 titles are common for User4 and User1.

Table 4.3. A Sample of Jaccard Similarity Values for Users.

User name 1	User name 2	Jaccard Similarity Value
User1	User2	0.102899907
User1	User3	0.209797657
User1	User5	0.065700483
User4	User1	0.105571848
User8	User1	0.099688474
User4	User5	0.046953047
User7	User9	0.096551724
User10	User3	0.1843044
User6	User4	0.028571429
User8	User7	0.119158879

Table 4.4. Distribution of Jaccard Similarity Values.

0%	25%	50%	75%	100%
0.00140647	0.03163017	0.04761905	0.06666667	0.25263158

Distribution of Jaccard Similarity result is given below in boxplot in Figure 4.5:

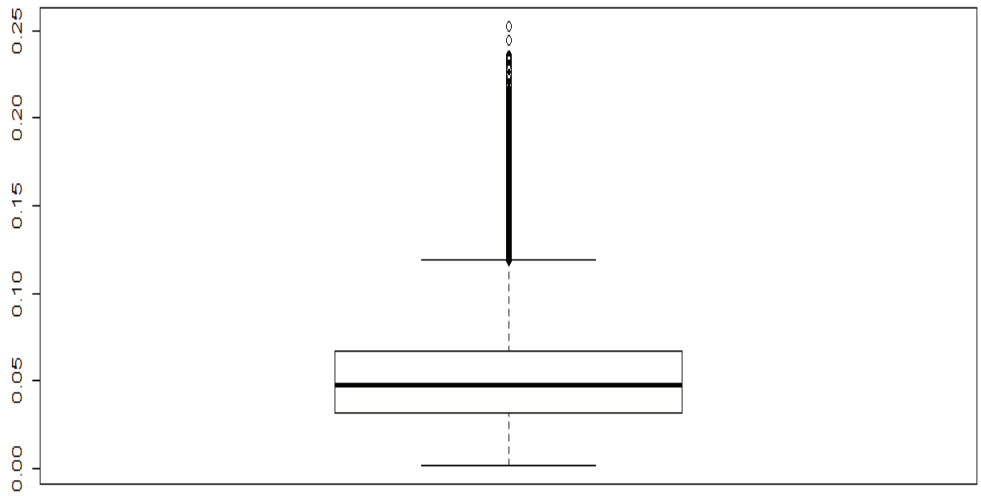


Figure 4.5. Boxplot of the Distribution of Jaccard Similarity Values.

In the light of the distribution of Jaccard similarity values, the interquartile range is between 0.031 and 0.066. We pruned user pairs which have similarity value less than

0.031 evaluating them as outliers. After that, the collaboration network is created. Lastly, the network file is converted into an input format ready to be processed by the BigCLAM algorithm.

4.3.2. Co-Occurrence Matrix and Word Association Network Creation Processes

Until this point, we crawled the data from Ekşisozluk and performed some natural language preprocessing operations on them. Here, we want to calculate the similarity of words in Turkish and then create a word association network.

In the first place, corpora was organized in titles. After that, word frequencies were calculated for each word and an index was assigned to all words. For instance, the most seen word got index zero. Then, data were read again title by title and a co-occurrence matrix was created. It has 10000 rows and 1000 columns.

After that, words were analyzed on specific window size. In this process, one word was selected analyzable word then, algorithm goes to forward for 10 words and goes to back for 10 words. These words were found on the co-occurrence matrix and their cell values were increased one. This operation was done for all words in the corpora and values were assigned on co-occurrence matrix.

Secondly, positive pointwise mutual information values were calculated for every word and these values were written into cells. For instance, positive pointwise mutual information of the word that is in second index and another word that is in fourth index were calculated and wrote into (2, 4) and (4, 2) cells. If the value was calculated less than zero, then zero was assigned into the cell. The new matrix was calculated with positive pointwise mutual information values.

Thirdly, cosine similarity was calculated for each word. Cosine similarity calculates the cosine angle of two vectors which are words in this scenario and if the result is close to one then these two words are very similar, if the result is close to zero then these two words are not similar. Sample table consisting of word pairs and their cosine similarity values are given below.

Table 4.5. A Sample Table for Cosine Similarity Values for Word Pairs

Index 1	Index 2	Cosine Similarity Result
444 - Siyasi	3083 - Siyasal	0.75810924
172 – Devlet	396 - Halk	0.56053099
8460 - TBMM	2296 - Meclis	0.65794965
8443 – Motive	8191 – Motivasyon	0.50334736
8421 – Tasarruf	9235 – Birikim	0.57344878

Lastly, the input was created for creating the word association network. Our social media data is very large and complex that's why only first 10000 words were considered in calculations. After that, BigCLAM input was created from the word association network. While creating word association network, pruning operation was done in similar logic with collaboration network creation process.

4.4. Overlapping Community Detection Algorithm Processes

Upto this point, we have one dataset from Ekşisözlük and we want to create two different networks from the same data. In this research, we have social media information and in social as well as other types of networks, nodes can be members of multiple communities simultaneously, which leads to overlapping community structures [28]. At this stage, BigCLAM, an overlapping community detection method is used for the creation of networks.

BigCLAM algorithm has some configuration to be completed before running. Configuration example is given below in Table 4.6.

Configuration descriptions are given below:

- Input edge list shows which nodes have relationships with each other and algorithm runs with this input file.
- Input file name for node names is used for analyzing network easily on other programs. BigCLAM reads this file and match ids from input edge list and tag labels one by one. For instance, labels show user names in a

collaboration network. In a word association network, labels show words.

- The number of communities to detect configuration is used for deciding the community number of the network.
- Minimum and maximum number of communities to try configuration is used for deciding the correct community number of the network. Therewithal, how many trials for the number of communities configuration has relationship with these two configurations. BigCLAM splits parts into trials number between minimum and maximum number of communities inputs and calculates likelihood.
- Number of threads for parallelization is used the for performance of the algorithm.

Table 4.6. BigCLAM Algorithm Configuration Example Table

Configuration	Input
Input edge list file name	forBigClam.txt
Input file name for node names (ID, Label)	nodeName.txt
The number of communities to detect (detect -1 automatically)	-1
Minimum number of communities to try	5
Maximum number of communities to try	100
How many trials for the number of the communities	10
Number of threads for parallelization	10

Firstly, collaboration network was studied using BigCLAM. In the first run, BigCLAM algorithm detects the number of communities automatically and we gave other input values like in Table 4.6. After that, BigCLAM calculates likelihood values and these values are given below in Table 4.7.

Table 4.7. Collaboration Network Number of Communities – Likelihood Values

Number of Communities	Likelihood Value
5	-10905.622170
6	-11015.433315
8	-11584.533811
10	-11221.723100
13	-11041.911118
17	-10812.258082
22	-11581.967549
29	-11738.237026
39	-11876.873764
52	-11528.101534
100	-11933.812029

Number of communities – likelihood graph is given below in Figure 4.6.

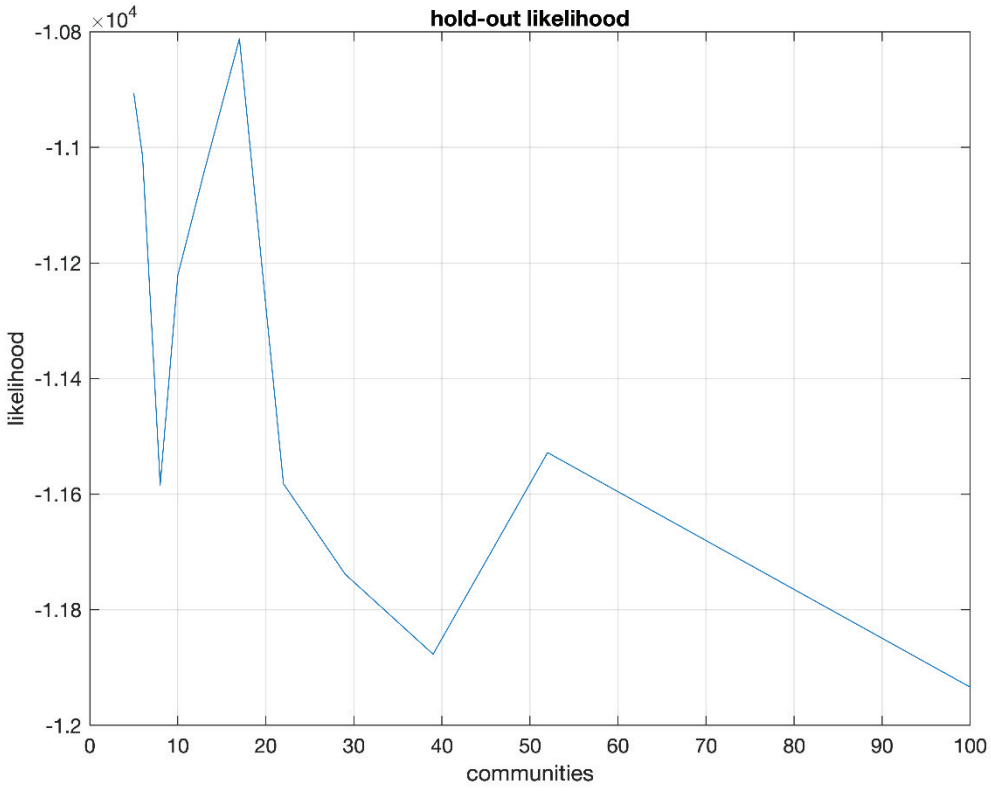


Figure 4.6. Collaboration Network Number of Communities – Likelihood Graph

BigCLAM algorithm calculated the likelihood values depending the number of communities of the collaboration network. The most important configuration to decide and configure is the community number of the network that's why community number of the network must be tuned correctly. Our aim was to detect the point at which the likelihood value does not change to a large extent. We used the elbow method for tuning community number both for collaboration network and word association network. Elbow method shows the minimal alteration point of the data. In the light of the likelihood values, we decided the community number of the collaboration network that is 29.

After that, BigCLAM algorithm was run again and configurations were tuned, and the community number of the collaboration network was calculated as 29 and the collaboration network was created.

Secondly, word association network was studied using BigCLAM. In the first run, BigCLAM algorithm detects the number of communities automatically and we gave other input values like in Table 4.6. BigCLAM algorithm calculated likelihood values which are given below in Table 4.8:

Table 4.8. Word Association Network the Number of Communities – Likelihood Values

Number of Communities	Likelihood Value
5	-4345469.980558
6	-4110099.095508
8	-4258834.230623
10	-4238548.887313
13	-6635852.484859
17	-7590387.465884
22	-8588870.002237
29	-9682388.136205
39	-7325656.575537
52	-10804580.138692
100	-17038546.429019

The number of communities – likelihood value graph is given below for the word association network in Figure 4.7.

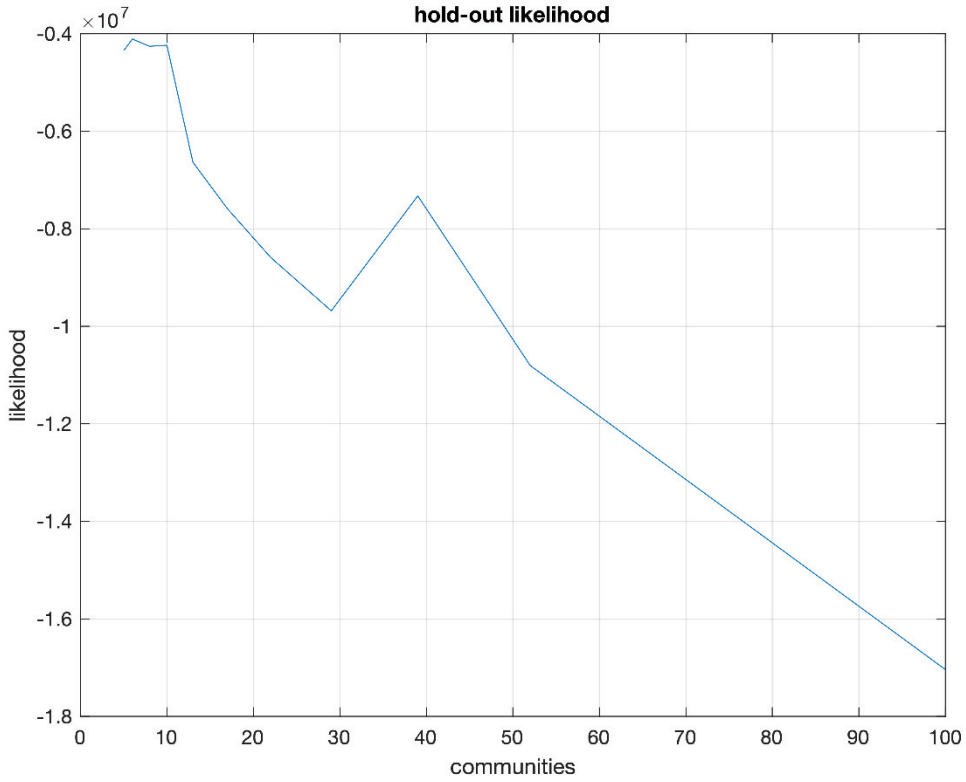


Figure 4.7. Word Association Network Number of Communities – Likelihood Graph

BigCLAM algorithm calculated likelihood values depending on the community number for the word association network. Again, our aim was to detect the point at which the likelihood value does not change to a large extent, the same perspective with collaboration network. After that, the elbow method was used to decide the number of communities. Elbow method shows the minimal alteration point of the data. In the light of the likelihood values, we decided the community number of the word association network that is 22.

After the first run, BigCLAM algorithm was run again with the correct number of communities and the word association network was created. Finally, two different networks were created with BigCLAM algorithm from the same social media data and these networks have overlapping community structures.

4.5. Network Analysis and Visualizations

Collaboration network and word association network are created. The collaboration network contains social media users and shows their similarities and the word association network contains words which are used by social media users and words associations. Before all, collaboration network is shown in Gephi with Force Atlas 2 layout below in Figure 4.8:

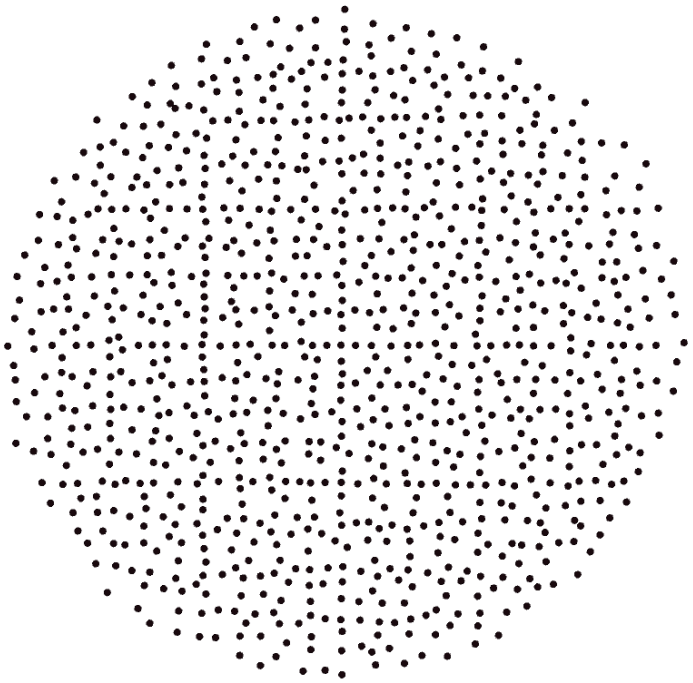


Figure 4.8. Collaboration Network

The collaboration network has 1000 users or nodes and 482.883 edges. The word association network is shown in Gephi with Force Atlas 2 layout below in Figure 4.9. The word association network has 10000 words or nodes and 4.487.849 edges. These two networks are created from the same social media data and we try to show the possible connections between the collaboration and the word association networks. In Ekşisözlük, users write entries and while writing they use words. We calculated user similarities while creating the collaboration network with Jaccard similarity and, we calculated word similarities while creating the word association network with Cosine similarity. Then we asked if users are similar then can these users use the similar words?

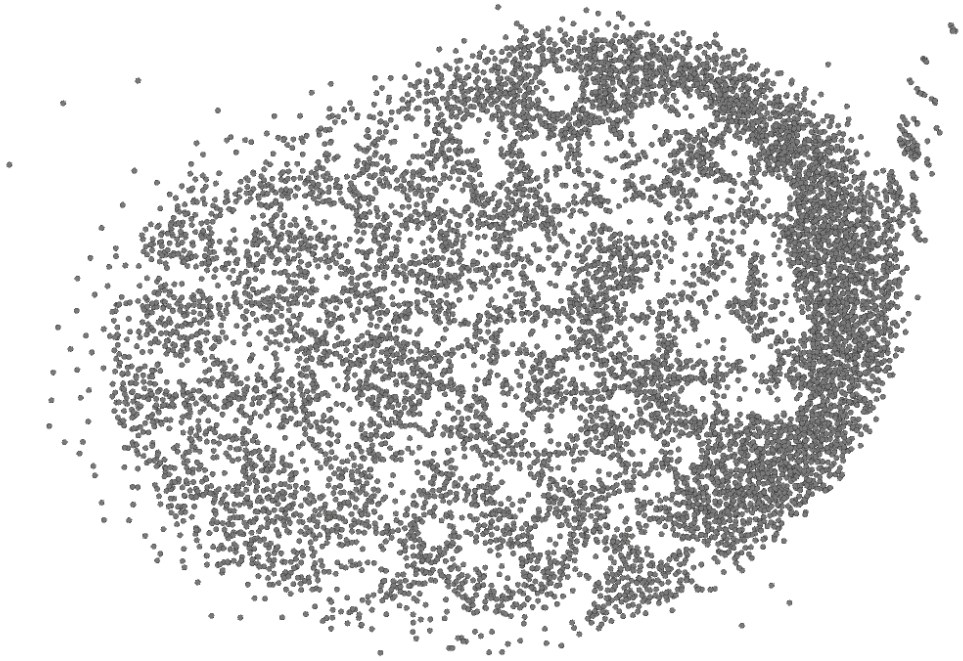


Figure 4.9. Word Association Network

Before the analysis of networks, we calculated for every user the top 30 words, which are put into entries. After that, we picked up some users whose similarity values are high from the collaboration network. We analyzed both the collaboration network and the word association network with these users and words and we found some connections between these two networks. Some result links are given below in Table 4.9.

Selected users' similarities are calculated with Jaccard similarity and given in Table 4.9 and these users are found in some similar communities in the collaboration network and set of words which are used by these users are found in some similar communities in the word association network. Some results are given below for the collaboration network in Table 4.10 and for the word association network in Table 4.11.

Table 4.10. shows that, User A and User B are members of 5 and User E and User C are members of 3 common communities. Table 4.11. shows that Bölüm and Dizi are members of 19 and Türkiye and Ülke are members of 20 common communities. In the light of these results, two different networks have some connections with each other. Similar users may use similar words. Also, similar users and similar words are in the same communities in their networks. These calculations support each other.

Table 4.9. Some Links Between the Collaboration and the Word Association Network

User name 1	User name 2	Jaccard Similarity Value	Word 1	Word 2	Cosine Similarity Value
User A	User B	0.128797084	Türkiye	Ülke	0.562382406
User C	User D	0.097847358	Bölüm	Dizi	0.822004022
User E	User C	0.063318777	Spoiler	Dizi	0.790048970
User E	User C	0.063318777	Güzel	Mükemmel	0.513951713
User F	User G	0.084615385	Futbol	Takım	0.846772698
User F	User G	0.084615385	Galatasaray	Futbol	0.850595944
User H	User I	0.058004640	Sene	Yıl	0.650911729
User J	User K	0.077142857	Uzun	Kısa	0.456982036
User L	User M	0.078740158	Siyasi	İktidar	0.696634705
User L	User M	0.078740158	Devlet	Siyasi	0.559439205

Table 4.10. User – Common Community Count in the Collaboration Network

User name 1	User name 2	Common Community Count
User A	User B	5
User E	User C	3
User F	User G	1
User H	User I	1

Table 4.11. Word – Common Community Count in the Word Association Network

Word 1	Word 2	Common Community Count
Bölüm	Dizi	19
Türkiye	Ülke	20
Güzel	Mükemmel	18
Sene	Yıl	22

After that, we looked for patterns of interests. Some specific communities in the collaboration network may match with some communities in the word association network. In this analysis, we found all users who are members of c4 in the collaboration network. Then, we calculated for every user in c4 the top 30-word list, which are put into entries. From the retrieved list, sports and film topics are observed as evidenced by the following words:

- Spoiler, Film, Dizi, Futbolcu, Sezon, Gol, Galatasaray

Community 4 has 287 users. In Table 4.12, the respective word – frequencies are given. Table 4.12 shows that, 77 of 287 users used the word spoiler in top 30-word list and this user is a member of community 4. Also, 37 of 287 users used the word film in top 30 words list. After that, we focused on the communities c0 and c25.

This time, particularly political terms are seen in the top 30-word list of users. In this process, we found members of c0 and c25. Then we analyzed 30-word list of these users and we saw that the words are related to politics, country's agenda. Also, these users did not use the words dizi, film, futbol and Galatasaray. In Table 4.13, some examples of users are given from c0 and c25 below.

Table 4.12. Word – Frequencies in Community 4 in the Collaboration Network

Word	Count
Spoiler	77/287
Film	37/287
Dizi	38/287
Futbolcu	12/287
Sezon	15/287
Gol	13/287
Galatasaray	11/287
Spoiler & Film	16/287
Spoiler & Dizi	27/287
Film & Dizi	14/287

Table 4.13. Examples from c0 and c25

User 1	Türkiye	User 3	Türkiye
User 1	Ülke	User 3	Ak
User 1	Oy	User 3	Chp
User 1	Seçim	User 3	Oy
User 2	Türkiye	User 4	Türkiye
User 2	Oy	User 4	Ülke
User 2	Terör	User 4	Türk
User 2	Pkk	User 4	Halk

Thus, Table 4.13 shows that, some mappings can be established between communities in the collaboration and word association networks as witnessed by c4 and c0, c25 communities in the collaboration network. The topics of sports and film can be assigned to c4 and politics can be associated with c0 and c25.

4.6. Comparison of Overlapping and Non-Overlapping Networks

Networks contain communities. Communities may have non-overlapping, sparse overlapping or dense overlapping structures. Real social media networks are so complex that they have dense overlaps [28]. We found a word association network that has dense overlaps. On the other side, if this network has had a non-overlapping community structure then these networks may be different by similarity.

Our real social media data are divided into communities with modularity [20]. "*Modularity has one issue that has received a considerable amount of attention is the detection and characterization of community structure in networks meaning the appearance of densely connected groups of vertices with only sparser connections between groups*" [20]. In the new generation algorithms, real social media networks are divided into communities which are overlapping that's why the network which is created with a non-overlapping structure and the network which is created with an overlapping structure could not be the same. While analyzing, we used modularity and BigCLAM and we compared the results.

Firstly, word association network was created again with modularity. The first network has 12 communities and the second network has 44 communities. The network which is created with BigCLAM has 22 communities and we created it before. The new community number of networks was selected specifically. One of them has communities nearly half of the first network and the other has communities twice bigger than the first network.

Normalized Mutual Information (NMI) is a good calculation procedure for determining the quality of clustering [25]. Since it is normalized, similarity can be measured and compared with NMI between different clusterings having different number of clusters. Also, rand index can be used for comparison operation for networks. First, these two networks were compared and NMI and Rand Index values were calculated.

First, network which has 12 communities was compared with BigCLAM communities. In the calculation phase, communities which are found by modularity are turned into matrices which have member count X 1. Then community members get specific numbers community by community. For instance, the first community members get 1, the second community members get 2. Details are given below in Table 4.14.

Table 4.14. Preparing 12 Modularity Communities for NMI Calculation

Community Name	Member Count	Specific Number
X0	10	1
X1	438	2
X2	2326	3
X3	21	4
X4	19	5
X5	10	6
X6	38	7
X7	9	8
X8	9	9
X9	5419	10
X10	760	11
X11	941	12

After that, all calculations turned into vectors. Then these vectors were added with each other and assigned a new variable that is a vector. Then communities which are found by BigCLAM are turned into matrices which have member count X 1. Then community members get specific numbers community by community, but this preparation has some different situations from the first one. In overlapping community structures:

- Some nodes may be shared, and these nodes may be members of a lot of communities.
- Some nodes may be unattended into any communities.

In the light of this information, before assigning a specific number, communities were sorted by member count. First operation was done on the community which has the biggest number of member count. After that, if shared nodes are found in every assign operation the category of these nodes was selected as the community which has smaller size. Details are given below in Table 4.15.

After that, all calculations turned into vectors. Then these vectors were added with each other and assigned a new variable that is a vector. Then, non-assigned nodes were determined from the last result and the specific number was selected and this number was not assigned before any community member. It is 23 for this analysis. After that, NMI value and rand index were calculated for the modularity result and BigCLAM result. The result values are given below in Table 4.16.

Results show that, clustering quality and similarity are low, and these two networks do not resemble with each other. This result is an expected result because social media networks are so complex, and non-overlapping community structure is inappropriate for social networks.

Secondly, network which has 44 communities was compared with BigCLAM communities. Calculation logic is same with calculation above. Communities which are found by modularity turns into matrices which have member count X 1 . Then community members get specific numbers community by community. Details are given in Table 5.1 which is found in appendix part.

Table 4.15. Preparing BigCLAM Communities for NMI Calculation

Community Name	Member Count	Specific Number
C0	604	1
C2	596	2
C1	593	3
C3	593	4
C4	591	5
C5	575	6
C6	569	7
C8	569	8
C10	563	9
C9	551	10
C7	548	11
C11	542	12
C16	541	13
C18	541	14
C12	537	15
C14	536	16
C13	535	17
C15	535	18
C20	529	19
C17	526	20
C19	520	21
C21	502	22

Table 4.16. Comparing Results for 12 – 22 Communities

Name	Result
Normalized Mutual Information	0.0386
Rand Index	0.4257

After that, these results were added with each other and assigned a new variable. Then same operation on first calculation was done for BigCLAM communities. Detail of specific number of members are given in Table 4.15. Again, if sharing nodes are found in every assign operation then category of these nodes was selected the community which have smaller size. Then, non-assigned nodes were determined from last result and the specific number was selected and this number was not assigned before any community member. It is 45 for this analysis. After that, NMI value and rand index was calculated for modularity result and BigCLAM result. The result values are given below in Table 4.17:

Table 4.17. Comparing Results for 44 – 22 Communities

Name	Result
Normalized Mutual Information	0.0383
Rand Index	0.3882

Results show that, clustering quality and similarity are low, and these two networks do not resemble with each other. This result is an expected result because social media data is inappropriate for non-overlapping community structure. This result is very close to first result. This situation shows that, if community count increases in folded then similarity results are very close.

Thirdly, non-assigned nodes may be important to calculate of the similarity rate. In this perspective, when comparing detected communities against ground truth, nodes are removed without ground-truth labels from the detected communities to achieve meaningful comparisons [5]. First, unassigned members were removed from dataset and network was created by modularity and it has 12 communities again. Then, every community members get specific numbers community by community like in Table 4.13 and Table 4.22. In this analyzing, we did not have any unassigned nodes that's why any assign operation is done. NMI and Rand Index values were calculated. The result values are given below in Table 4.18. The summary table shows that, removing unassigned members increases similarity results.

Results show that, clustering quality and similarity are low but removing operation of unassigned members increases results. Normalized mutual information

value is better than nearly 5 times from first calculation. Rand index value is better than first calculation. Summary table is given below in Table 4.19.

Table 4.18. Comparing Results for 12 – 22 Communities Removing Unassigned

Name	Result
Normalized Mutual Information	0.1520
Rand Index	0.4567

Table 4.19. Summary Table for 12 – 22 Communities

Name	Remove Unassigned	Result
Normalized Mutual Information	No	0.0386
	Yes	0.1520
Rand Index	No	0.4257
	Yes	0.4567

After that, new network was created by modularity and it has 44 communities. While creating the new network, unassigned members were removed from dataset. Then, every community members get specific numbers community by community like in Table 4.15 and Table 5.1. In this analysis, we did not have any unassigned nodes that's why any assign operation is done. NMI and Rand Index values were calculated. The result values are given below in Table 4.20:

Table 4.20. Comparing Results for 44 – 22 Communities Removing Unassigned

Name	Result
Normalized Mutual Information	0.1576
Rand Index	0.4614

Results show that, clustering quality and similarity are low again, but removing operation of unassigned members increases results as expected. Normalized mutual information value is better than nearly 5 times from first calculation. Rand index value is better than first calculation. Summary table is given below in Table 4.21.

Table 4.21. Summary Table for 44 – 22 Communities

Name	Remove Unassigned	Result
Normalized Mutual Information	No	0.0383
	Yes	0.1576
Rand Index	No	0.3882
	Yes	0.4614

The summary table shows that, removing unassigned members increases similarity results.

Alternatively, qualities of detected communities can be measured in a different way. "*Although a global best match (i.e., finding a one-to-one mapping) between detected communities and ground-truth communities would be ideal. We used per community best match as a heuristic alternative*" [5]. Specifically, the research defines the average F_1 score [5].

$$F_1 = \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k} max_j\, F_1\left(A_i, B_j\right) + \frac{1}{k'}\sum_{j=1}^{k'} max_i\, F_1(B_j, A_i)\right)$$

First analysis was done with network, which has 12 communities and BigCLAM communities. Unassigned members were removed from dataset before, then communities were created by modularity. Every community turns into matrices which have member count X 1. If a community contains a member in specific indices, value of this indices assigns 1. This operation was done because matrix dimensions must same for F_1 score calculation.

After that, modularity communities and BigCLAM communities were compared with each other. For instance, F_1 score value of X0 community and C0 community is found then F_1 score value of X0 and C1 communities is found. C0, C1, C2, C3, …, C21 are calculated for X0 and the biggest F_1 score value is found for X0. Same operation is done for X1, X2, X3, …, X11 and all biggest values are found.

Then, BigCLAM communities and modularity communities were compared with each other. For instance, F_1 score value of C0 community and X0 community is found then F_1 score value of C0 value and X1 communities is found. X0, X1, X2, …, X11 are calculated for C0 and the biggest F_1 score value is found for C0. Same operation is done for C1, C2, C3, …, C21 and all biggest values are found. Then these values use into formula. The result value is given below in Table 4.22:

Table 4.22. F_1 Score Result for 12 – 22 Communities

Name	Result
F_1 Score	0.4141

The F_1 score shows that, clustering quality and similarity are low as expected. Because social media networks are so complex, and social media dataset is appropriate with overlapping community structure and non-overlapping community structure gives different results from result of overlapping community structure.

Second analysis was done with network which has 44 communities and BigCLAM communities. Unassigned members were removed from dataset before, then communities were created by modularity. Other operations are same with previous calculation. Modularity communities and BigCLAM communities were compared with each other. For instance, F_1 score value of X0 community and C0 community is found then F_1 score value of X0 and C1 communities is found. C0, C1, C2, C3, …, C21 are calculated for X0 and the biggest F_1 score value is found for X0. Same operation is done for X1, X2, X3, …, X43 and all biggest values are found.

Then, BigCLAM communities and modularity communities were compared with each other. For instance, F_1 score value of C0 community and X0 community is found then F_1 score value of C0 value and X1 communities is found. X0, X1, X2, …, X43 are calculated for C0 and the biggest F_1 score value is found for C0. Same operation is done for C1, C2, C3, …, C21 and all biggest values are found. Then these values use into formula. The result value is given below in Table 4.23:

Table 4.23. F_1 Score Result for 44 – 22 Communities

Name	Result
F_1 Score	0.4129

The F_1 score shows that, clustering quality and similarity are low as expected if the community number is changed. Because social media dataset is appropriate with overlapping community structure and non-overlapping community structure gives different results from result of overlapping community structure.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Social media is a very large and complex environment to analyze. It is an artifact and contains intertwined objects. In this thesis, we studied on social media data. The social media data were crawled and stored in a database. The language of the collected data is in Turkish. Before all, we studied on Turkish natural language processing. Zemberek library was integrated into the environment. Many NLP operations were applied on social media data like tokenization, stop word removal and stemming.

As part of the research, we created two different types of network namely collaboration and word association networks from the same social media data and identified communities on them. We created collaboration network with users and we created word association network with words. In collaboration network, user similarities were calculated with Jaccard similarity and similar users were found. In word association network, co-occurrence matrix was created, and word similarity was calculated with positive pointwise mutual information and cosine similarity. We aimed to show some connections between users in the collaboration network and vocabulary sets in word association networks. We calculated the top 20 words which are written for every user. After that, we linked up some users and words with similarity calculations. We linked up some connections in collaboration and word association networks.

Overlapping community detection is a useful tool for analyzing complex structures like social networks. Before that, non-overlapping community structure was used. Social media is a very complex environment and social media data in Turkish was analyzed with state-of-the-art BigCLAM overlapping community structure algorithm. Also, non-overlapping community structures were created from the same social media data with modularity. The network which was created with modularity and the network which was created with BigCLAM were compared with each other. The claim of overlapping community algorithms is that the non-overlapping algorithms are inappropriate to complex environments. These two networks were compared by clustering evaluation metrics and results show that they do not resemble each other.

In future, mappings between networks may be based on some rules. Communities of the word association network may be associated with topics like sports, politics or art. Also, the collaboration network may show hobbies and area of interests of users.

REFERENCES

1. Affiliation Networks / Bipartite Networks http://www.shizukalab.com/toolkits/sna/bipartite (accessed Jan 05, 2019).

2. Chinchor N., In MUC-4 evaluation metrics, *Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics*: **1992**; pp 22-29.

3. Deshpande, M.; Karypis, G. J.; Item-Based Top-N Recommendation Algorithms, *ACM Transactions on Information Systems*, **2004**, 22 (1), 143-177.

4. Devi, J. C.; Poovammal, E., An analysis of overlapping community detection algorithms in social networks, *Procedia Computer Science*, **2016**, 89, 349 - 358.

5. Du, R.; Kuang, D.; Drake, B.; Park, H., Hierarchical Community Detection via rank-2 Symmetric Nonnegative Matrix Factorization, *Computational Social Networks*, **2017**, 4 (1), 7.

6. Durugbo, C.; Hutabarat, W.; Tiwari, A.; Alcock, J. R., Modelling Collaboration using Complex Networks, *Information Sciences*, **2011**, 181 (15), 3143-3161.

7. Eustace, J.; Wang, X.; Cui, Y., Community Detection using Local Neighborhood in Complex Networks, *Physica A: Statistical Mechanics; Applications*, **2015**, 436, 665-677.

8. Girvan, M.; Newman, M., Community Structure in Social and Biological Networks, *Proceedings of the National Academy of Sciences*, **2002**, 99 (12), 7821-7826.

9. Guan, J.; Yan, Y.; Zhang, J. J., The Impact of Collaboration and Knowledge Networks on Citations, *Journal of Informatics*, **2017**, 11 (2), 407-422.

10. Huang, H., Design, Analysis and Experimental Evaluation of a Distributed Community Detection Algorithm, Master Thesis, **2015**.

11. Hubert, L.; Arabie, P., Comparing Partitions, *Journal of Classification*, **1985**, 2 (1), 193-218.

12. Jaccard, P., Étude comparative de la distribution florale dans une portion des Alpes et des Jura. **1901**, 37, 547-579.

13. Kaimal, R., Document Summarization using Positive Pointwise Mutual Information, **2012**.

14. Kowalski, G. *Information Retrieval Systems. Theory and Implementation, 1st edn (Printed in the USA; Kluwer Academic Publishers);* ISBN 0-7923-9899-8: 1997.

15. Lancia, F. J.; Word Co-occurrence and Similarity in Meaning, Mind as Infinite Dimensionality Charlotte, NC: *Information Age Publishers*, **2007**.

16. Li, Y.; Luo, P.; Wu, C., A New Network Node Similarity Measure Method and Its Applications, **2014**.

17. McDaid, A. F.; Greene, D.; Hurley, N., Normalized Mutual Information to Evaluate Overlapping Community Finding Algorithms, **2011**.

18. Model-based Approach to Detecting Densely Overlapping Communities in Networks. http://snap.stanford.edu/agm/ (accessed Apr 15, 2018).

19. Mogotsi, I.; Manning, C.D.; Raghavan, P.; Schütze, H., Introduction to Information Retrieval. Springer: 2010.

20. Newman, M., Modularity and Community Structure in Networks, *Proceedings of the National Academy of Sciences*, **2006**, 103 (23), 8577-8582.

21. Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. J., Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. **2005**, 435 (7043), 814.

22. Peel, L.; Larremore, D. B.; Clauset, A., The Ground Truth about Metadata and Community Detection in Networks, *Science Advances*, **2017**, 3 (5), e1602548.

23. Rajendra, Q. W.; Raj, J. D.; Recommending News Articles using Cosine Similarity Function, *Warwick Business School Journal*, **2015**, 1-8.

24. Shahaf, D.; Guestrin, C.; Horvitz, E.; Leskovec, J., Information Cartography, *Communications of the ACM*, **2015**, 58 (11), 62-73.

25. Strehl, A.; Ghosh, J., Cluster Ensembles---A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, **2002**, 3 (Dec), 583-617.

26. Tokenization.https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html (accessed Apr 12, 2018).

27. Turney, P. D.; Pantel, P., From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, **2010**, 37, 141-188.

28. Yang, J.; Leskovec, J., In *Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach*, Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM: 2013; pp 587-596.

29. Zhou, X.; Liu, Y.; Wang, J.; Li, C., A Density Based Link Clustering Algorithm for Overlapping Community Detection in Networks, *Physica A: Statistical Mechanics; Applications*, **2017**, 486, 65-78.

APPENDIX A

MODULARITY

Table 5.1. Preparing 44 modularity communities for NMI calculation

Community Name	Member Count	Specific Number
X0	646	1
X1	16	2
X2	304	3
X3	8	4
X4	28	5
X5	1969	6
X6	8	7
X7	8	8
X8	8	9
X9	8	10
X10	16	11
X11	8	12
X12	8	13
X13	16	14
X14	150	15
X15	7	16
X16	8	17
X17	8	18
X18	7	19
X19	7	20
X20	7	21
X21	8	22
X22	8	23

(**cont. on next page**)

Table 5.1 (cont.)

Community Name	Member Count	Specific Number
X23	8	24
X24	692	25
X25	8	26
X26	53	27
X27	8	28
X28	769	29
X29	8	30
X30	8	31
X31	5073	32
X32	8	33
X33	8	34
X34	8	35
X35	15	36
X36	8	37
X37	8	38
X38	16	39
X39	8	40
X40	8	41
X41	8	42
X42	9	43
X43	8	44