



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**  
**2020-2021**

| PROJECT TITLE                                       |
|---|
| Smart Discussion Platform                           |
| FACULTY ADVISOR                                     |
| Prof. Dr. Reda ALHAJJ                               |
| TEAM MEMBERS  |
| 64160002-Mert Can ÇAKMAK<br>63170005-Bedirhan GERİN |



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

**Project Code**

**Project Title:**

Smart Discussion Platform

**Faculty Advisor:**

Reda ALHAJJ

**Project Team Members:** Mert Can ÇAKMAK, Bedirhan GERİN

**Sponsor Company (if any) :**

| BUDGET (TL)             | PROPOSED | CONSENTED |
|-------------------------|----------|-----------|
| IMU FUNDING             | 82150 TL | 82150 TL  |
| SPONSOR COMPANY FUNDING | -        | -         |
| TOTAL                   | 82150 TL | 82150 TL  |

| PROJECT PLAN  | PROPOSED   | CONSENTED  |
|---------------|------------|------------|
| PROJECT PLAN  | 05.10.2020 | 05.10.2020 |
| STARTING DATE | 05.10.2020 | 05.10.2020 |



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

| ADVISOR   | DEPARTMENT CHAIR  |
|---|---|
| <b>Name:</b> Reda ALHAJJ  | <b>Name:</b> Mehmet Kemal ÖZDEMİR   |
| <b>Contact Information:</b><br>Tel :<br>E-mail : ralhajj@medipol.edu.tr | <b>Contact Information:</b><br>Tel :<br>E-mail : mkozdemir@medipol.edu.tr |
| <b>Signature:</b>   | <b>Signature:</b>   |

| TEAM MEMBER   | TEAM MEMBER   |
|---|---|
| <b>Name:</b> Bedirhan GERGİN  | <b>Name:</b> Mert Can ÇAKMAK  |
| <b>Contact Information:</b><br>Tel : 505 064 21 10<br>E-mail : bedirhangergin@gmail.com | <b>Contact Information:</b><br>Tel : 538 445 82 20<br>E-mail : mccakmak@st.medipol.edu.tr |
| <b>Signature:</b>   | <b>Signature:</b>   |



# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

**Project Title:** Smart Discussion Platform

**Faculty Advisor:** Prof. Dr. Reda ALHAJJ

**Team Members:** Mert Can ÇAKMAK, Bedirhan GERİN

**Project Group Title:**

#### ABSTRACT

Microblogs and social media platforms are spreading rapidly. Within these platforms, people are expressing their feelings and opinions about a specific situation or topic. This creates a big flow of data. So, in this project, it is aimed to develop a Smart Discussion Platform that allows people to discuss topics and also allows administrators to rule and analyze the system in terms of trends, influence, trust issues by ML(Machine Learning) and NLP(Natural Language Processing) Techniques.

On the website, people will create accounts with their personnel information. They will comment on any topic they want. Also, the platform will have some features to indicate the reactions of other people's comments such as the like and dislike buttons. On the other side, the platform will be aiming to analyze all the inputs. Firstly, the popularity of the topics will be analyzed, trending topics will be determined. Secondly, comments on the topics will be analyzed. Positive and negative comments will be evaluated. Sentiment analysis will be done to analyze the context of comments. Thirdly, according to comments and reactions, trust and influence issues will be analyzed through the users. Also, another important target will be about invitation policy. Social media such as Twitter will be analyzed. According to analysis, an invitation will be sent to every influencer people who have commented on the relative topics to argue this topic in our platform.

In the project, mainly three algorithm development is aimed for the functions stated above along with the website creation. Fundamental of all these algorithms were data collection and preprocessing. After these methods are identified, the first experiment and methods were implemented for the sentiment analysis algorithm. After the creation of the algorithm, validation of the algorithm was tested for the preliminary result. Further development will be done by creating other algorithms and website in the future steps.

**Keywords:** microblogs, natural language processing, data preprocessing, sentiment analysis, trust and influence



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

**Project Code**

**BUDGET PROPOSED– (TL)**

|                             | ITEMS    |                    |           |         |        |
|-----------------------------|----------|--------------------|-----------|---------|--------|
|                             | PEOPLE   | MACHINE-INSTRUMENT | MATERIALS | SERVICE | TRAVEL |
| <b>IMU FUND</b>             | 72000 TL | 10000 TL           | 0         | 150 TL  | 0      |
| <b>SPONSOR COMPANY FUND</b> | -        | -                  | -         | -       | -      |
| <b>TOTAL</b>                | 72000 TL | 10000 TL           | 0         | 150 TL  | 0      |

**BUDGET APPROVED– (TL)**

|                             | ITEMS    |                     |            |         |        |
|-----------------------------|----------|---------------------|------------|---------|--------|
|                             | PEOPLE   | MACHINE-INSTRUMENT* | MATERIALS* | SERVICE | TRAVEL |
| <b>IMU FUND</b>             | 72000 TL | 10000 TL            | 0          | 150 TL  | 0      |
| <b>SPONSOR COMPANY FUND</b> | -        | -                   | -          | -       | -      |
| <b>TOTAL</b>                | 72000 TL | 10000 TL            | 0          | 150 TL  | 0      |

- 2 People, 4500 TL month/per person, 8-month total.
- 2 Computer, 5000 TL per computer.
- Website Domain per year 20 \$



# **Istanbul Medipol University**

## **School of Engineering and Natural Sciences**

### **Graduation Project**

#### **1. PROJECT PURPOSE**

Developing a Smart Discussion Platform that allows people to discuss topics and also allows administrators to rule and analyze the system in terms of trends, influence, trust issues by Machine Learning and Natural Language Processing Techniques.

#### **2. PROJECT TARGET**

Project targets to create a website as a platform that allows people to discuss any topic and subject. People will create accounts with their personnel information. They will comment on any topic they want. Also, the platform will have some features to indicate the reactions of the other people's comments such as like and dislike buttons.

On the other side, the platform will be aiming to analyze all the inputs.

Firstly, the popularity of the topics will be analyzed, trending topics will be determined. Secondly, comments on the topics will be analyzed. Positive and negative comments will be evaluated. Sentiment analysis will be done to analyze the context of comments.

Thirdly, according to comments and reactions, trust and influence issues will be analyzed through the users.

Also, another important target will be about invitation policy. Social media such as Twitter will be analyzed. According to analysis, an invitation will be sent to every influencer people who have commented on the relative topics to argue this topic on our platform.

#### **3. SCOPE**

The kind of platform that people share their ideas spreading rapidly. Within these platforms, people are expressing their feelings and opinions about a specific situation or topic. This creates a big flow of data. So, we want to create our platform and have our data to process. We want to provide a smooth environment for the people to express their ideas, whereas we also aim to benefit from these opinions by evaluating and analyzing them.

When we look at the analyzing part, our first scope border will be analyzing the data by performing linguistic analysis and explaining the discovered phenomena from this collection. This will be possible by building a sentiment classifier which will enable us to determine positive, negative, and neutral sentiments for the comments of people.

Sentiment is an opinion or feeling about some subject or person. Sentiment analysis finds out the emotion, opinion or feeling in the context. To be exact, choosing whether a particular discussion is positive, negative, or unbiased. Sentiment analysis has wide applications such as emotion mining, opinion mining etc.[1]. As more and more user comments about topics they argue or express their view, our platform will be more valuable sources for people's opinions.

Another important point will be processing the raw data. Before the sentiment and other analyses, data should be ready. The first step of this will be extracting the stop words such as "the", "a", "with" etc. Words like this do not contain any emotion. When we throw out our emotionless words, the analysis can proceed to another level. The rest of the words will be classified according to the determined structure. This structure can be extracted from an already written dictionary, program, and library that will score the words as their attitudes toward the topics. These words will be graded and labeled as positive, negative, and neutral.

It is also important to have a mechanism that will assess the reliability of users and their influences to provide a useful online discussion platform. There are various techniques to determine the influence of a person. Some of the works assign trust values to users to indicate

# **Istanbul Medipol University**

## **School of Engineering and Natural Sciences**

### **Graduation Project**

their trustworthy degrees according to a structured principle like the interaction of the people's comments and posts. We will create our system and structure for this duty. Some of the available algorithms also will be taken into consideration. One of them is the PageRank algorithm. It was created by Google to rank web pages. This algorithm measures the relative significance of a page inside a chart by filtering the complete web. There are lots of algorithms that take the PageRank as their fundamental idea and convert it for the people and social media. We will also benefit from this kind of algorithms and create a hybrid model.

Lastly, social media such as Twitter will be analyzed. According to analysis, an invitation will be sent to every influencer people who have commented on the relative topics to argue this topic on our platform. This will also require the development of a structure to determine the people and the topics from this platform. We will use a similar structure to determine the sentiment analysis and influences on these platforms.

#### **4. SUCCESS CRITERIA**

Designing a good database to get high performance and fast response from the website. Forming a good looking and not a complex interface for the website. Making the website practical to use. For this purpose, survey will be done through website to understand the user satisfactory. Aim will be getting at least %85 from this survey.

Designing an accurate sentiment analysis algorithm, to extract sentences that are in favor of the subject topic or sentences are that against the subject topic.

Designing accurate trust and influence analysis algorithms to determine the user's trust factor and influence factor.

Inviting most trustable and influencer people to the platform from microblog sites and at the same time inviting people that are sharing opinions that are in favor of the subject topic or against the subject topic according to the course of the discussion.

For the algorithms at least %85 accuracy will aimed to get the success criteria.

#### **5. ORIGINALITY and ADDED VALUE**

Data that people poured into the internet like reactions and comments on the topics have the potential to uncover precious experiences on human feelings. Thus, the analysis of people's ideas and comments can play a crucial role to get people's behavior and response in various ways. With the increasing number of microblogs and social media, individuals have started to specific their conclusions on a wide assortment of themes on Twitter and other comparable stages. As they are growing and spreading rapidly these tools becoming more useful to understand and model various events. Enlightening of these, we aim to create our platform that values this fact as a fundamental principle.

Some other platforms aim to provide a discussion environment for their users with their regulation and policies. Whereas Twitter is focusing more on the people to express their feelings and opinions with a given limit of words. But the common thing is that every one of them focuses on just providing the necessary environment for the people's feelings, opinions in a structured way. We also want to create an environment for the people that they can freely and comfortably express their feeling about any topic. Additionally, as a fundamental principle, we will analyze the system by ML and NLP Techniques to enlighten trends, influence, and trust issues.

# **Istanbul Medipol University**

## **School of Engineering and Natural Sciences**

### **Graduation Project**

#### **6. LITERATURE REVIEW**

There are works about sentimental analysis, measuring the influence and the trust of the user, and topic modeling.

In the Sentiment Analysis and Influence Tracking using Twitter paper [1], the authors mention that how Twitter information is utilized as a corpus for investigation by the application of assumption investigation and a think about of distinctive calculations and strategies that offer assistance to track the impact and effect of a specific user/trademark dynamic on the social arrange. They used Twitter API, Twitter Streaming API, Twitter Search API for data collection. For analysis preprocessing, techniques are used which are tokenization, normalization, and part of speech (POS) tagging. To determine the influence of the user PeopleRank and TwitterRank algorithms are used. Using these data collection APIs data can be collected from Twitter easily and ranking algorithms can help to calculate the influence of the user.

In the Detecting Real-World Influence Through Twitter paper [2] the authors investigated the issue of identifying the real-life impact of individuals based on their Twitter account. For the dataset CLEF RepLab, 2014 dataset is used. SNA(Social Network Analysis), PCA(Principal Component Analysis), bag of words, POS, linear classifiers which are Support Vector Machine (SVM) and libLinear, logistic regression, logic boost, multinomial Naïve Bayes are used for determining real-world influence. Since bots are not real influence in the real world this is helpful to detect someone's real influence value.

In the Topic Modeling of Twitter Conversations paper [3], the creators displayed a way to analyze expansive sums of literary information from Twitter discussions effectively and viably. Particularly, it was clarified how to capture the stories that individuals share on Twitter around social occasions, decrease their complexity, and give conceivable clarifications. For this Latent Dirichlet Allocation (LDA) method is used. By using this method, the topics from contexts can be extracted efficiently and effectively.

In the Extracting health-related causality from Twitter messages using natural language processing paper [4], the authors assessed an approach to extracting causalities from tweets using natural language processing (NLP) techniques. Twitter Streaming API is used for dataset collection. To extract causality lexicon syntactic relations and NLP pipeline operations which are lemmatizing, POS and dependency parsing is used. Since a good causality relationship sentence results in the good influence of a person when a reader reads that sentence so that this can be used for determining the influence of the user. In any case, there are numerous diverse ways to state cause and impact relations in the sentence it is hard to hold all the sets.

In the Investigating the Relationship between Trust and Sentiment Agreement in Arab Twitter Users paper [5] the creators proposed a investigate technique system for exploring the relationship between trust and sentiment understanding on Twitter, and clarify the system by applying it to a utilize case from Saudi Arabia. For this, the adaptation of the EigenTrust Algorithm which is the MarkovTrust algorithm is used. Also, surface analysis, deep analysis, and shallow analysis algorithms are used to determine the relationship between trust and sentiment agreement. Since the context and sentiment have taken into consideration determining the trust of the user will be more accurate.

In the Influence Analysis of Emotional Behaviors and User Relationships Based on Twitter Data paper [6], the authors analyzed the impact of enthusiastic behaviors on client



# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

connections based on Twitter information utilizing two lexicons of passionate words. For the collection of data random sampling, for calculation emotion score Keyword Matching, and the testing Brunner-Munzel test is used. By looking at emotional behaviors the influence of the user can be determined.

To sum up and show related works, below chart was prepared.

| Author   | Period | Title   | Method/Remarks   |
|--|--------|---|--|
| Pramila M. Chawan  | 2012   | Sentiment Analysis and Influence Tracking using Twitter                                     | “Paper makes a sentiment analysis on an entity level; mine people's idea on specific entities instead of whole tweets, scrutinize each tweet. Use three main features for scoring: followers, mentions and retweets, and lists, these are used for ratio to users” [1]   |
| Peiyao Li , Weiliang Zhao, Jian Yang, and Jia Wu   | 2019   | CoTrRank: Trust Evaluation of Users and Tweets  | “Develop a trust ranking approach named CoTrRank. It mainly uses a coupled dual network. By evaluating the coupling affect in users and tweets. Values are derived with their original meanings in different trust spaces. The results of experiment show that the CoTrRank provides better evaluations of the trustworthiness of users and tweets when it is compared with other methods.” [7]  |
| Jean-Valère Cossu, Nicolas Dugué, Vincent Labatut  | 2015   | Detecting Real-World Influence Through Twitter  | “Project analyze Twitter-based features with comparing and allowing to measure the offline effects and influence of users. Look for specific characteristics in twitter that can explain individual known to be prominent in their real-life.” [2]   |
| Son Doan , Elly W. Yang, Sameer S. Tilak, Peter W. Li, Daniel S. Zisook and Manabu Torii | 2018   | Extracting health-related causality from twitter messages using natural language processing | “Causality extraction is done by outputs that are dependency parser of Lexico-syntactic patterns. These techniques were used to help and improve the preciseness of information extraction. Paper shows that dependency parser with lexicon-syntactic relations yields high precision, which is an important feature for big data set mining.” [4]   |
| Kiichi Tago and Qun Jin  | 2018   | Influence Analysis of Emotional Behaviors and User Relationships Based on Twitter Data      | “Paper conduct three different experiments: calculate the mean feeling score of a client, calculate the mean feeling score utilizing passionate tweets, and calculate the mean feeling score utilizing enthusiastic tweets, with not including users of few emotional tweets. Then analyze by Brunner–Munzel test for the influence of emotional behaviors to user relationships. From the result it is understand that a positive user is more operative than a negative user for building a user relationship in a specific situations.” [6] |
| Areeb Alowisheq and Sarah O Al-Humoud  | 2017   | Investigating the Relationship Between Trust and Sentiment Agreement in Arab Twitter Users  | “It conducts a research method for identifying the relationship between trust and sentiment for Arab Twitter users.” [5]   |

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

|  |      |   |  |
|--|------|---|--|
| Youngguae Bae and Hongchul Lee   | 2012 | Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers | “Paper identify between the positive and negative group of onlookers of popular tweet users. Then, find that the audience are influenced by the sentiments used in the tweets by popular users. Thirdly, from these two findings it develops a positive-negative measurement for influence. Finally, by a Granger causality analysis, it is understood that sentiment alter of the audience was related to the real- world sentiment scene of well known clients.” [8] |
| Cano Basave, A. E.; Mazumdar, S. and Ciravegna, F.                     | 2011 | Social influence analysis in microblogging platforms a topic sensitive based approach                       | “Paper suggests the use of lexical profiles forming dominant users depending upon the retweet Twitter graph. Establishes a different version of the PageRank algorithm for examining user’s relevance of a retweet connection.” [9]  |
| Juyup Sung, Seunghyeon Moon, and Jae-Gil Lee                           | 2013 | The Influence in Twitter: Are They Really Influenced?   | “Paper tenders a development of PageRank algorithm, which is InterRank. It regards both relationship and topical similarity among users. It suggests that topical similarity act upon dominance.” [10]   |
| Eliana Sanandres , Camilo Madariaga , Raimundo Abello                  | 2018 | Topic Modeling of Twitter Conversations   | “Paper suggests a technique for topic modeling on Twitter chatting which is Latent Dirichlet Allocation to decide the topics that are talked.” [3]   |
| Liangjie Hong and Brian D. Davison                                     | 2010 | Empirical Study of Topic Modeling in Twitter  | “Paper suggests a solution for normal topic model algorithms that have been used on social media. It proposes that training a topic model with clustered text, it can be achieved better accuracy and preferable performance.” [11]  |
| Christan Grant, Clint P. George, Chris Jenneisch, and Joseph N. Wilson | 2011 | Online Topic Modeling for Real-time Twitter Search  | “Paper aims to get the attractive and topical social media entries from the dataset. It uses topic modeling algorithm for examination in the dataset.” [12]  |
| Elias Jonsson, Jake Stolee   | 2016 | An Evaluation of Topic Modelling Techniques for Twitter   | “Paper assesses of different topic modelling algorithms and analyze them by looking their performance on Twitter texts.” [13]  |
| Yefeng Ruana, Arjan Duresia, Lina Alfantoukha                          | 2018 | Using Twitter Trust Network for Stock Market Analysis   | “Paper suggests that using the trust between users on microblogs, this can improve the mutual affinity with budgetary information within the stock advertise.” [14]  |

*Table 1. Related Works*

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

## 7. EXPERIMENTS/METHODS

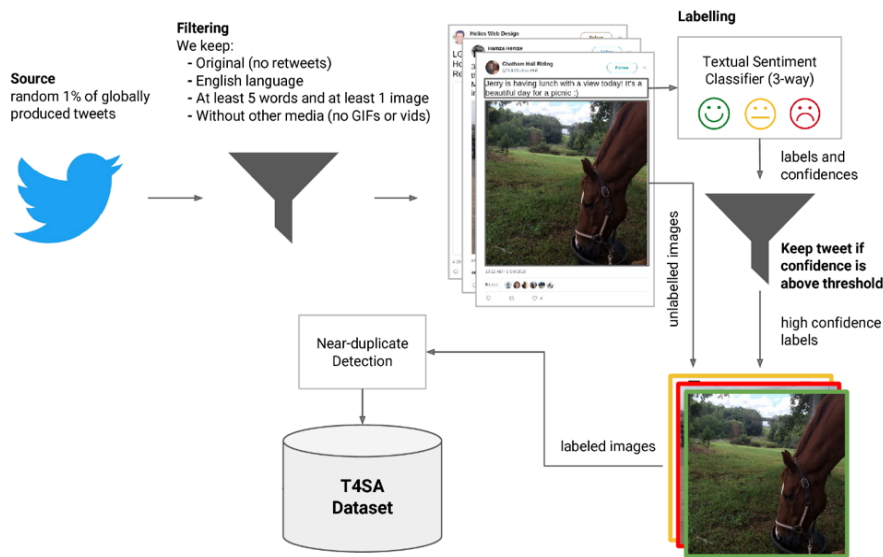
### 1) Sentiment Analysis Algorithm

On the way of project aims, first experiment and methods were implemented for the sentiment analysis algorithm.

First of all, a training set was implemented to build up a model. On the other hand, after creating and modeling the algorithm, a test (or validation) dataset was used to validate the model. The training dataset is consisting of two different datasets that are taken directly from different sources. One is T4SA Dataset and the other one is Sentiment140 Dataset.

#### T4SA Dataset:

T4SA Dataset consist of tweets that are collected with Twitter Sample API from July to December 2016 around 6 months. Also, tweets were gone through some evaluation by discarding some of them which are not containing any static image or other media, not written in English, retweets and whose text was less than 5 words. In the project the images were not used, only tweet texts were used for sentiment classification. So, also images and text were separated in the process. [15]



*Image 1. T4SA Dataset*

After these processes, dataset contains approximately 3.4 Million tweets. Each tweet was labeled according to the sentiment in the tweets predicted by LSTM-SVM architecture. The dataset includes three sentiments which are positive, neutral and negative.

The dataset had scores between 0 and 1 which indicates the sentiment of the text. For confidence and not to lose much of the data, 0.85 threshold was determined to take the tweets. However, it was seen that this dataset was not enough for our model. Because positive and especially negative data were insufficient when it is compared with neutral data. To solve this problem the dataset was merged with Sentiment140 dataset.

#### Sentiment140 Dataset:

Sentiment140 Dataset was created by looking emoticons. Smile faces have label as positive, sad faces have label as negative. The Twitter Search API was used to collect the

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

tweets and the Maximum Entropy classifier was used to classify the sentiment. [16]  
It consist of 821191 negative, 460857 neutral and 1067747 positive tweets with total 1067747 tweets.

| sentiment | label | count   |
|-----------|-------|---------|
| negative  | -1.0  | 821191  |
| neutral   | 0.0   | 460857  |
| positive  | 1.0   | 1067747 |

Table 2. Imbalanced Data

After dataset is examined, it was seen that data was imbalanced. So data was undersampled to make equal distributions with the number of tweets. Result can be seen in the table below:

| sentiment | label | Count  |
|-----------|-------|--------|
| negative  | -1.0  | 458700 |
| neutral   | 0.0   | 460857 |
| positive  | 1.0   | 459111 |

Table 3. Balanced Data

After acquiring and preparing the training dataset, preprocessing techniques were started for preparing the raw data for further processing. Preprocessing steps are important steps since it transforms text into a more digestible and understandable form so that machine learning algorithms can perform better.

#### Preprocessing

The preprocessing steps are:

- 1) **Lower Tweets:** Text are transformed to lowercase.
- 2) **Remove the URLs:** Links that have "http" or "https" or "www" are removed by empty string.
- 3) **Remove mentions, retweet and hashtags:** Words starting with "@", "#", "RT" are removed
- 4) **Remove symbols:** Emoticons, symbols and pictographs, transport and map symbols, flags, other language characters and dingbats are removed.
- 5) **Remove non alphabet characters:** Characters except Digits and Alphabets with a space are removed
- 6) **Remove consecutive letters three or more:** Consecutive letters are replaced by just their 2 letters. (eg: "Cooool" to "Cool")
- 7) **Remove punctuations:** Punctuations are removed from the sentence since it is not affecting the meaning of the sentence.
- 8) **Remove stopwords:** The stopwords the words that are does not add meaning to a sentence.



# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

After preparing the last version of the datasets, model training was started.

#### **Training the Data**

First of all  $X_{train}$  and  $X_{test}$  datasets are transformed into matrix of TF-IDF features by using the TF-IDF Vectoriser. This dataset will be used to train the model and test against it. Also The  $y_{train}$  and  $y_{test}$  datasets indicates the label of the sentence (-1, 0, 1).

Since the boundary conditions were determined with training the data, the next task was to predict the target class with the classification. To get more accurate result and compare it with other results, multiple classification methods were used.

#### **Classification methods**

##### **Naïve Bayes**

Naive Bayes is a classification technique that uses the Bayes theorem with an independence of between predictors as an assumption.

##### **Multinomial Naive Bayes**

Multinomial Naive Bayes algorithm counts for multiple features that occurs and it will classify a dataset according to counts it finds of multiple keywords.

##### **Bernoulli Naive Bayes**

Bernoulli Naive Bayes counts the single feature occurrence and counts for the same feature that does not exist. It focus just on a single keyword, also it counts how many times that keyword does not exist in the document.

##### **Support Vector Machine**

Support Vector Machine algorithm analyzes the data for classification and regression analysis. It outputs the map of sorted data and their margins between these two as different as possible.

##### **Linear Support Vector Classification**

Linear SVC is similar to SVC, it supports only a linear kernel. It has more flexibility in the choice of parameters Also, it is faster and can scale a lot better than SVC.

##### **Multinomial Logistic Regression**

The logistic regression is used to guess the probability of a certain class or event existing such as positive/negative sentences. Also, the logistic regression can be used to guess several classes of events such as positive/neutral/negative sentences. Logistic Regression is used when the dependent target is categorical.

#### **Data Collection**

Implementing the sentiment algorithm and using for further steps in the project also data collection method is identified. Collecting the data from a social media website was done through a scraper. A scraper is a type of software used to copy content from a website. In this project Snsrape was used for this purpose. Snsrape is a tool for social networks. It is used for scraping things like user profiles, hashtags, or searches and returns them.

Below is an example data collection that were taken from Twitter and transformed into csv file:



# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

|    | A                   | B                | C   |
|----|---------------------|------------------|---|
| 1  | date                | user             | tweet   |
| 2  | 2020-12-25 23:59:58 | Wfdeell1         | @LLinWood @TXPSALM55 You got that right I think this whole covid-19 deal was a plan from China and the demotard's to take out Trump but it backfired on them they didn't think Trump would handle it as well as |
| 3  | 2020-12-25 23:59:53 | KennaStevens1    | @TheRickyDavila @doxiedachsie When they were informed of COVID19 early this year they invested in a company that makes body bags. Why would they care about the nations virus death toll? These two are de      |
| 4  | 2020-12-25 23:59:52 | JenMangler       | Short but important thread. #COVID19 #COVID #edchat #aedchat https://t.co/85Z2nVUMQB  |
| 5  | 2020-12-25 23:59:51 | COEmergency      | COVID-19 vaccine administered: 63,170 doses #COVID19Colorado https://t.co/1ArRHVva4 https://t.co/fmW2ptXQYA   |
| 6  | 2020-12-25 23:59:51 | AntiTotalitaBot  | @CTVNews inaccurate counting: How COVID-19 Deaths Are Counted https://t.co/gjMyeINtV2   |
| 7  | 2020-12-25 23:59:46 | TestUser05632971 | Some Passengers Infected After Man Died of COVID-19 on Plane https://t.co/GWXB20h5Ho  |
| 8  | 2020-12-25 23:59:45 | DutchDL          | Guerilla Mask Force Protest Denmark and Germany Covid-19. https://t.co/F1dv5bZjd via @YouTube   |
| 9  | 2020-12-25 23:59:43 | newsnov9ja       | News of COVID-19 Vaccine Special Gift of Christmas&#x201c;NLCA President https://t.co/t6kbA0ae4j  |
| 10 | 2020-12-25 23:59:41 | JaimeAnaya       | Suspensions grow that nanoparticles in Pfizer&#x201c;s COVID-19 vaccine trigger rare allergic reactions https://t.co/OB62aOxkUa   |
| 11 | 2020-12-25 23:59:41 | adeoke_mukaila   | The spread circumstances substance of covid-19, goes extremely breakout viral in the world. But we have the cause of glorified God @ last. compliment of the season,  |
| 12 | 2020-12-25 23:59:41 | EricKNG          | Covid-19 UK Mutant Strain: Higher Hospitalizations, Deaths Likely, Study Finds - Bloomberg https://t.co/3bfmmkYEL   |
| 13 | 2020-12-25 23:59:38 | julesofmaine     | @portlandimibier @nirav_maineccdc @MEPublicHealth The real truth about Covid-19. Respiratory deaths are not higher due to covid-19! We&#x201c;re being played. Wake up people! Amazing interview with Dr. Zach  |
| 14 | 2020-12-25 23:59:35 | govlossss        | @JimmyMac2021 @Toronto1880 @fordnation And Peel didnt?  |
| 15 | 2020-12-25 23:59:33 | jtmonrad         | And several @NIH researchers published this piece on "COVID-19 vaccine trial ethics once we have efficacious vaccines."   |
| 16 | 2020-12-25 23:59:33 | MartyKoekemoer   | @AFranzen @Thomas_Binder @tngadd We&#x201c;re becoming preoccupied with Covid19. And Death. And not seeing the damage that is Being done to the Living. And future generations. Fearful of being Shamed - fc    |
| 17 | 2020-12-25 23:59:27 | LiterateLiberal  | Yuma Prison Warden Dies From COVID-19 After Dismissing Safety Concerns   @crooksandlars https://t.co/bABIQ0bc8Y   |
| 18 | 2020-12-25 23:59:25 | myraluv2015      | Yes until covid19 is under control. https://t.co/7qalds0BR  |
| 19 | 2020-12-25 23:59:23 | PrincetonBoy915  | @Sharlie528 Covid-19? I don&#x201c;t know her &#x201c;8Y&#x201c;8Y&#x201c;  |
| 20 | 2020-12-25 23:59:22 | MartyKoekemoer   | @LusyNote @Thomas_Binder We&#x201c;re becoming preoccupied with Covid19. And Death. And not seeing the damage that is Being done to the Living. And future generations. Fearful of being Shamed - forgetting to |
| 21 | 2020-12-25 23:59:21 | jtmonrad         | Do these concerns apply here?   |
| 22 | 2020-12-25 23:59:18 | NewsThaivisa     | Two seafood vendors in Onnut fresh market test positive for Covid - https://t.co/dbBaNx06Hu   |

Image 2. Example Covid-19 Data from Snsrape

## 2)Website Construction

After Sentiment Analysis algorithm website was constructed.

Name of the website is Smart Discussion Platform. Main page is seen as below.

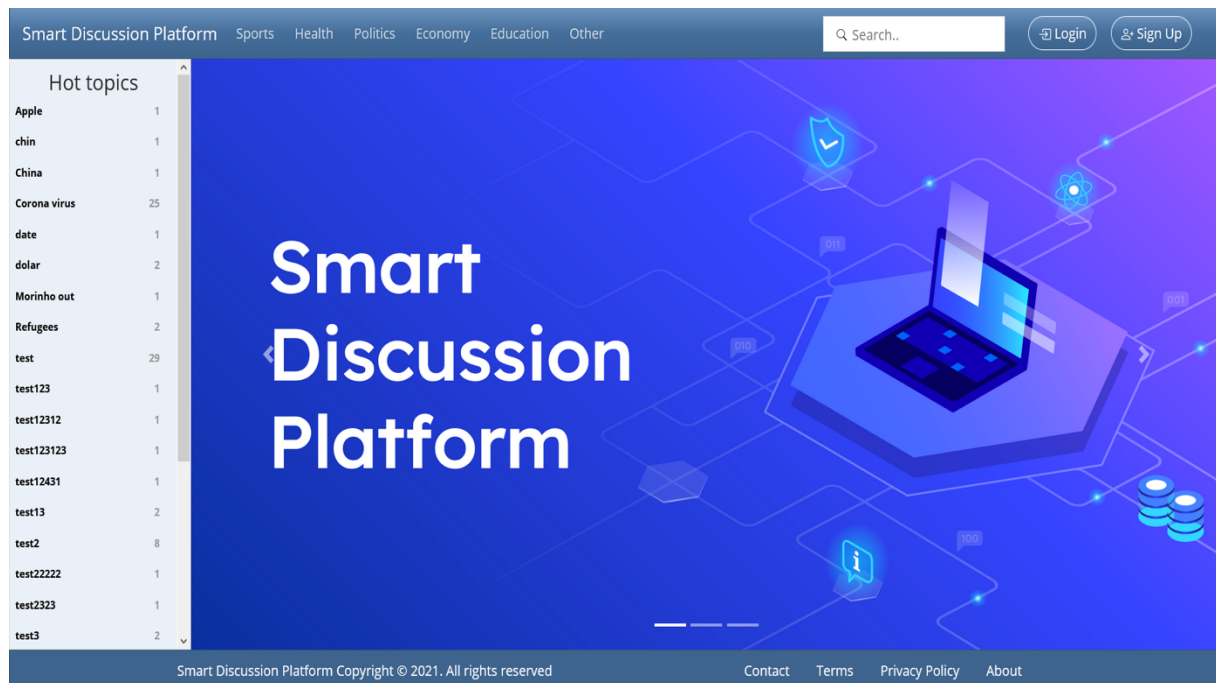


Image 3. Website Main Page

This is the page that welcome people when they entered to website. As it is seen, people can choose to login if they are already signed up or can sign up. Even if they don't signed up they can see the hot topics and can look at the topics as they wishes. But they can't comment or be privileged for any other functionalities since they are not signed up. After users are signed up the main page is seen as below.

# Istanbul Medipol University

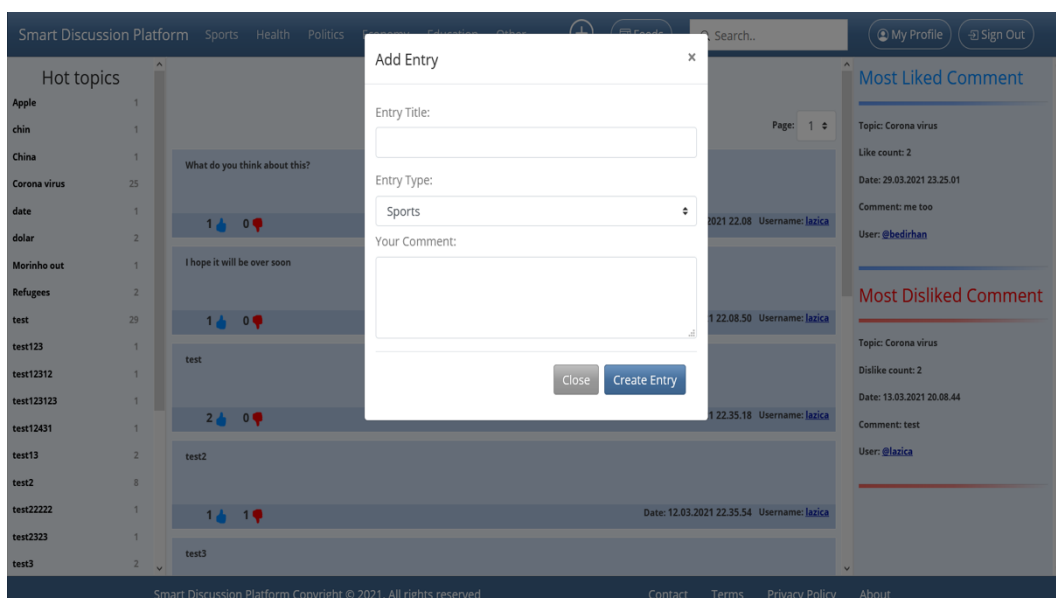
## School of Engineering and Natural Sciences

### Graduation Project



*Image 4. Website Main Page(2)*

From this main page they can enter the topics from the left hot topic list or they can enter according to subjects from the top navbar. After they enter the topics, they can like or dislike the comments under the topic or they can left their own comments to the topic. On the right side of the page they can also see the most liked and disliked comments in any topic. Also from the top navbar plus button they can open up their own topic to discuss as seen below:



*Image 5. Topic Adding*

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

From the main page they can direct to their own profile as it is seen below.

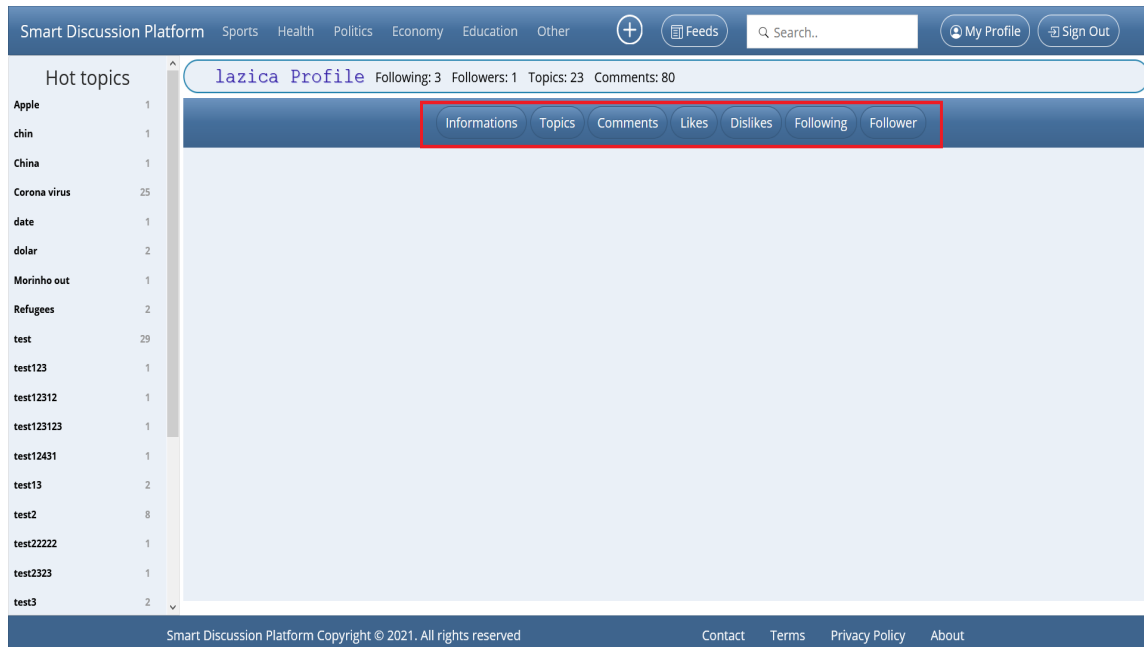


Image 6. Profile Page

From the profile page users can see their information as given in the picture such as followers, follows, comments and topic numbers. Also, they can change their personal information related to their profile. Also users can enter other people's profile pages as seen in the below.

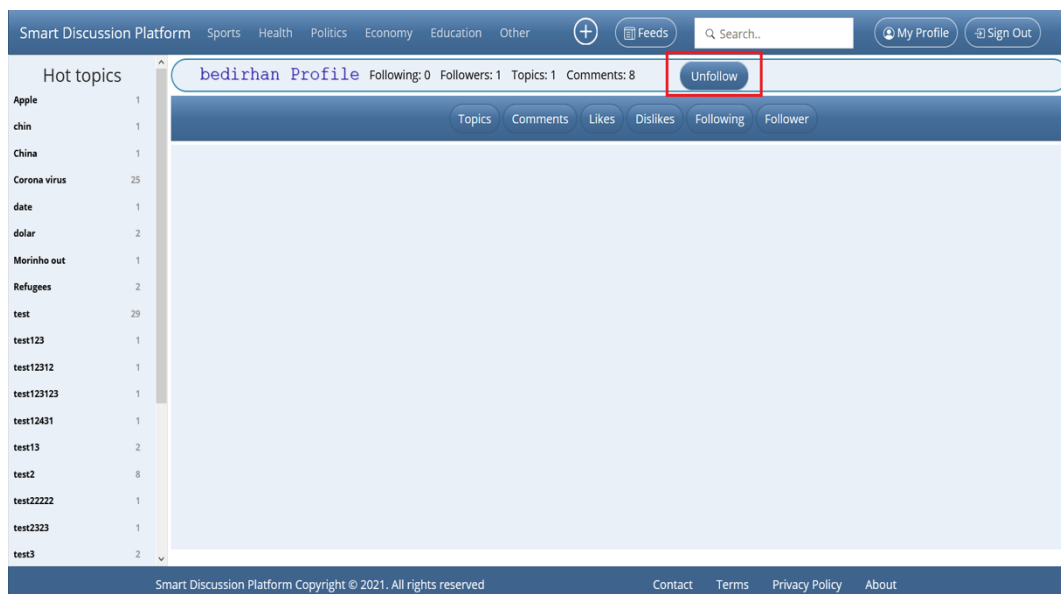


Image 7. Profile Page(2)



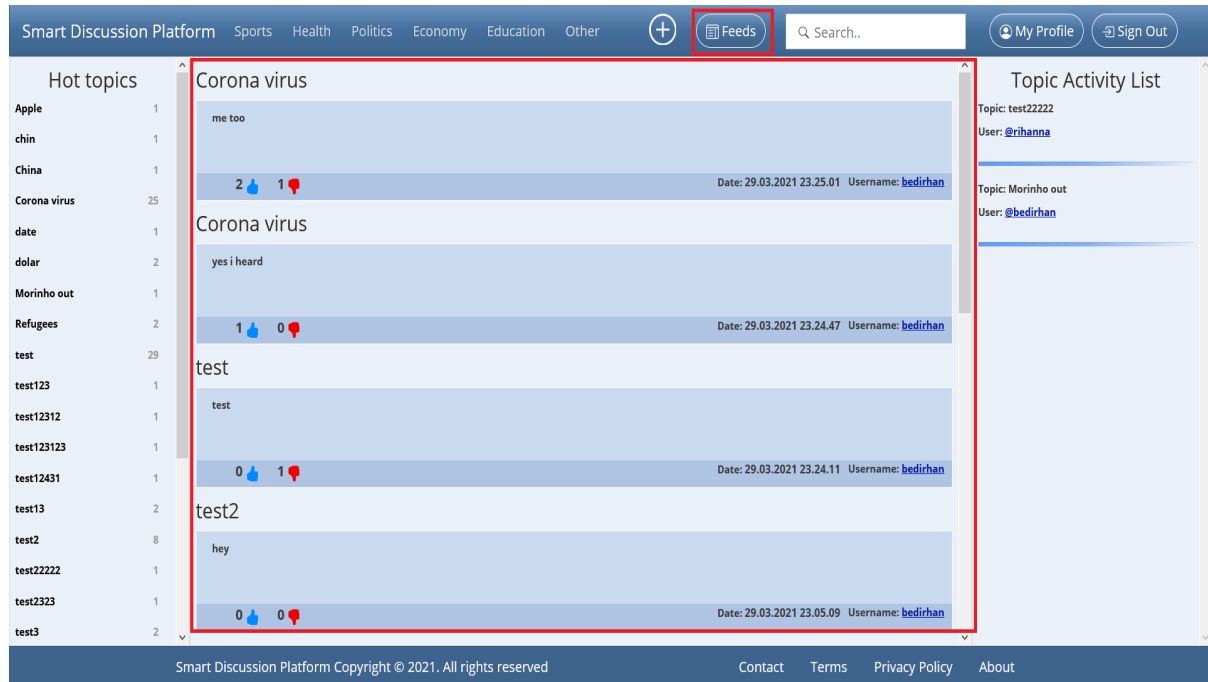
# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

From this page people can see the other people's information and also people can follow and unfollow the users from the follow button.

Lastly, with the Feeds button people are directed to a special page name Feeds where a timeline that shows the user activities that only a user follow occurs as seen below.



*Image 8. Feed Page*

### 3)Influence Algorithm

Our platform apply a social-networking “Following” model like twitter. In this model each user is allowed to follow who she/he wants to follow without any granting. Also, this user can also be followed by others without taking any permission first.

In most of the influence algorithm, just number of followers of a user is taken into consideration while analyzing. But with TwitterRank [17] algorithm, topic similarity is also introduced into the analyze. In our algorithm we also resemble and try to integrate the TwitterRank algorithm into our own platform.

It is revealed that in this kind of model, based on Twitter, and relationship between users, 72.4% of the users follow more than 80% of their followers, and 80.5% of the user have 80% of their friends follow them back [18]. This phenomenon is called “reciprocity”. Reciprocity can be explained in two option: First, a person can follow other person just randomly and this person can also follow back just for a good gesture of this relation. But on the other hand, a person follows a user because this user publishes or write about topics that he or she is interested in, and this user follows back because he or she thinks they share similar interest on topic. This relationship is called “homophily”, this phenomena is seen in many social networks [19]. Also, TwitterRank reveals that:

- i) Users with a very high probability, users with “following” relationships are more

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

similar than those without having any resemblance on the topics they are interested in.

- ii) Users with reciprocal “following” relation are tend be more similar than the users without according to the topics they are interested in.

This is a very good indicator for our purpose. With this approach while influence between users are analyzed, topic similarity is also counted since if a topic similarity exist between a two users who follows each other, has higher influence and has a meaningful relationship that can contribute to their influences.

So, in our algorithm users influence for each follower is calculated by the relative amount of content ,comment in our case, the follower received from another user. This similarity measure come from the one of the fundamental influence algorithm “PageRank” in measuring influence.

Also, since the influence on each user is based on exposure amount of content the follower receives, these contents are differ from user to user. They are less interesting even when the relative content have a big amount. Since users generally have different interests in various topics, influence of users also varies in different topics. Because of that, a topic-sensitive algorithm will be used. So, every user will have influence in the case of their topic interest.

On the way of algorithm, first a directed graph is created to understand the following relation between the users. Vertex set is  $V$ , which is for all the users. Edge set is  $E$ . There will be an directed edge between two users if user follows other user.

| User      | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-----------|----------|--------|-----------|---------|--------|---------|---------|
| bedirhan  | 1        | 0      | 0         | 0       | 0      | 0       | 0       |
| senior    | 1        | 1      | 0         | 0       | 1      | 0       | 0       |
| mert12345 | 0        | 0      | 1         | 0       | 0      | 0       | 0       |
| test123   | 0        | 0      | 0         | 1       | 0      | 0       | 0       |
| lazica    | 1        | 1      | 0         | 0       | 1      | 1       | 1       |
| rihanna   | 0        | 0      | 0         | 0       | 1      | 1       | 0       |
| mert123   | 0        | 0      | 0         | 0       | 0      | 0       | 1       |

*Table 4. Follower-User Matrix*

Then, to understand the every user’s topic interest a Topic interest matrix ( $T$ ) is created. In the platform every entry has a topic category. So if a comment is done under this entry, it means that this comment is under the topic of that category. So  $T$ , a  $U \times T$  matrix, where  $U$  is user’s numbers and number of topics is  $T$ .  $T_{ij}$  contains the number of times a comment in users has been assigned to topic  $t$ .

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

| User/Topic | Economy | Other | Politics | Health | Sports |
|------------|---------|-------|----------|--------|--------|
| bedirhan   | 0       | 3     | 0        | 3      | 2      |
| senior     | 3       | 8     | 2        | 2      | 4      |
| mert12345  | 0       | 0     | 0        | 0      | 0      |
| test123    | 0       | 0     | 0        | 0      | 0      |
| lazica     | 7       | 36    | 5        | 23     | 23     |
| rihanna    | 0       | 0     | 0        | 0      | 1      |
| mert123    | 0       | 0     | 0        | 0      | 0      |

*Table 5. User-Topic Type Relationship Matrix*

After that we row normalize the matrix such that  $\|T'_i\|=1$  for each row of  $T$ . Then now, each row of matrix  $DT$  shows the probability distribution of user interest in topic  $T$ . With the help of this probability matrix we will be able to find the topic similarity between users.

| User/Topic | Economy | Other | Politics | Health | Sports |
|------------|---------|-------|----------|--------|--------|
| bedirhan   | 0       | 0.375 | 0        | 0.375  | 0.25   |
| senior     | 0.158   | 0.421 | 0.105    | 0.105  | 0.210  |
| mert12345  | 0       | 0     | 0        | 0      | 0      |
| test123    | 0       | 0     | 0        | 0      | 0      |
| lazica     | 0.074   | 0.383 | 0.053    | 0.245  | 0.245  |
| rihanna    | 0       | 0     | 0        | 0      | 1      |
| mert123    | 0       | 0     | 0        | 0      | 0      |

*Table 6. Normalized User-Topic Type Relationship Matrix*

After creating these two matrices, we are ready to find influence scores of every users to their followers in a specific topic:

$$P_t(i,j) = \frac{|T_j|}{\sum_{a: s_i \text{ follows } s_a} |T_a|} sim_t(i,j)$$

$|T_j|$  is number of comments commented by  $s_j$  sums up the number of comments published by all of  $s_i$ 's friends. Similarity is also defined as:

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}|$$

From this notation for the first part we can infer that a user  $s_i$  follows a number of user and those users can write different numbers of comments, all of which will be directly visible to  $s_i$ . The more a user  $s_j$  comment, the higher portion of comments  $s_i$  expose from  $s_j$ . Generally, this will lead us to get a higher influence on  $s_j$

Second part is about topical similarity between users as proposed by the homophily.  $DT'$  shows the probability of users  $s_j$ 's interest in different topics. The similarity between two user such as  $s_i$  and  $s_j$  in a topic  $t$  can be calculated as the difference between the probability of two users who are interested in that topic  $t$ . The more similar the two users are, the higher the

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

influence score.

| User      | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-----------|----------|--------|-----------|---------|--------|---------|---------|
| bedirhan  | 1        | 0.842  | 1         | 1       | 0.926  | 1       | 1       |
| senior    | 0.842    | 1      | 0.842     | 0.842   | 0.917  | 0.842   | 0.842   |
| mert12345 | 1        | 0.842  | 1         | 1       | 0.926  | 1       | 1       |
| test123   | 1        | 0.842  | 1         | 1       | 0.926  | 1       | 1       |
| lazica    | 0.926    | 0.917  | 0.926     | 0.926   | 1      | 0.926   | 0.926   |
| rihanna   | 1        | 0.842  | 1         | 1       | 0.926  | 1       | 1       |
| mert123   | 1        | 0.842  | 1         | 1       | 0.926  | 1       | 1       |

*Table 7. User Similarity Matrix*

After finding every person's influence score to another user in a specific topic we are column normalizing to find the total influence number of a user in this topic which will also be able to find us the Influence Rank on that topic.

| User/Topic | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|------------|----------|--------|-----------|---------|--------|---------|---------|
| Economy    | 0        | 0.917  | 0         | 0       | 1.842  | 0       | 0       |
| Other      | 0.344    | 0.700  | 0         | 0       | 1.505  | 0       | 0       |
| Politics   | 0        | 0.948  | 0         | 0       | 1.895  | 0       | 0       |
| Health     | 0.606    | 0.344  | 0         | 0       | 1.517  | 0       | 0       |
| Sports     | 0.361    | 0.552  | 0         | 0       | 1.133  | 0.035   | 0       |

*Table 9. Influence Rank Matrix*

After summing up influence of a user in every topic, that will give us the total Influence Score of users.

| User            | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-----------------|----------|--------|-----------|---------|--------|---------|---------|
| Total Influence | 1.311    | 3.460  | 0         | 0       | 7.892  | 0.035   | 0       |

*Table 10. Total Influence Rank Matrix*

#### 4) Trust Algorithm

Trust is one of the fundamental criteria for human relation. There are many definitions to explain and define trust. For our case it is a measure of confidence that an entity or user will behave in expected manner. There are lots of application and model to represent the trust in the systems and every system has its own grading features to calculate trust. In many social media platform and social networking sites there are different kind of approaches for grading users trust towards another member.

One of the important algorithm in this manner is "Eigen Trust". It cores the network-based approach for calculating trust among peers. This algorithm is used in P2P networks and it shows subjective criteria because it uses users subjective ratings of their interaction with one another to compute trust.

In our project we also take the "Eigen Trust" algorithm as base and try to build up accordingly. There are different kind of interaction in our platform, one of the important one is "like" and "dislike" reaction. Users are expressing their outcomes to comments of other

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

user with this feature. So in our algorithm this feature is taken as resembling of weighting protocol of one user to another.

First of all, like in the influence algorithm, a directed graph is created to understand the following relation between the users. Vertex set is  $V$ , which is for all the users. Edge set is  $E$ . There will be an directed edge between two users if user follows other user. After that a matrix is created to see the every user's like numbers to create the like matrix  $U$ . So  $U$ , a  $U \times U$  matrix, where  $U$  is the number of users.  $U_{ij}$  shows the number of times user  $i$  in users has like user  $j$ 's comments. This relation is created if there is a follow relation between this two users.

As likes are important in this relation, also dislike are counted into consideration as negative impact on trust. But dislike are weighted less according to likes and number of dislikes are multiplied with weight 0.25. So as a result, matrix is finalized with trust evaluation scores ( $U_{ij}$ ) of a user to another user is calculated as:

$$U_{ij} = \text{likes} + (0.25)\text{dislikes}$$

| User      | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-----------|----------|--------|-----------|---------|--------|---------|---------|
| bedirhan  | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| senior    | 0        | 0      | 0         | 0       | 7      | 0       | 0       |
| mert12345 | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| test123   | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| lazica    | 3.75     | 0.75   | 0         | 0       | 0      | 1       | 0       |
| rihanna   | 0        | 0      | 0         | 0       | 2      | 0       | 0       |
| mert123   | 0        | 0      | 0         | 0       | 0      | 0       | 0       |

*Table 11. Trust Matrix*

After that we row normalize this matrix such that  $\|U'_{i,}\|=1$  for each row  $U$ . Each row of matrix  $U$  is basically the probability distribution of user trust over his following people. Also after normalizing that we have taken into consideration another parameter "Influence". We thought that the more a person is influencer the more his trust is important. So we took the influence score and scale them down to between 1 and 2. And when we are assigning the last version of Trust Matrix, we multiply the trust scores with people's influence score which is between 1 and 2. This means most influencer user's evaluation of another user will be multiplied and will be contribute two times of this score. On the other hand, least influencer people will have the factor 1 and his contribution value will be multiplied with 1. So the last version of trust matrix will be achieved.

| User      | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-----------|----------|--------|-----------|---------|--------|---------|---------|
| bedirhan  | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| senior    | 0        | 0      | 0         | 0       | 1.438  | 0       | 0       |
| mert12345 | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| test123   | 0        | 0      | 0         | 0       | 0      | 0       | 0       |
| lazica    | 1.364    | 0.273  | 0         | 0       | 0      | 0.364   | 0       |
| rihanna   | 0        | 0      | 0         | 0       | 1.004  | 0       | 0       |
| mert123   | 0        | 0      | 0         | 0       | 0      | 0       | 0       |

*Table 12. Normalized Trust Matrix*

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

After that matrix will be summed columns based. So every assigned trust value to that user will be summed up. As a result of this, Trust score of that user will be revealed which will enable us to rank the users among them.

| User  | bedirhan | senior | mert12345 | test123 | lazica | rihanna | mert123 |
|-------|----------|--------|-----------|---------|--------|---------|---------|
| Score | 1.364    | 0.273  | 0         | 0       | 2.443  | 0.364   | 0       |

*Table 13. Total Trust Matrix*

## 8. RESULTS

All the methods and implementation explain in the experiment part will be used to analyze the data that will be gathered from our platform. Result of this analysis will be shared in a special admin page in the platform.

First sentiment analysis algorithm;

According to methods implemented for sentiment analysis, results were gathered. Classification method results are as stated below:

| Method  | Accuracy           |
|---|--------------------|
| Multinomial Naive Bayes Training              | 0.818151332458568  |
| Multinomial Naive Bayes Test                  | 0.7919137856212932 |
| Bernoulli Naive Bayes Training                | 0.812743470675391  |
| Bernoulli Naive Bayes Test                    | 0.7923477506147837 |
| Linear Support Vector Classification Training | 0.8662902155838766 |
| Linear Support Vector Classification Test     | 0.8180963402285549 |
| Logistic Regression Training                  | 0.8204264042760245 |
| Logistic Regression Test                      | 0.8123101403153479 |

*Table 14. Methods and their Accuracies*

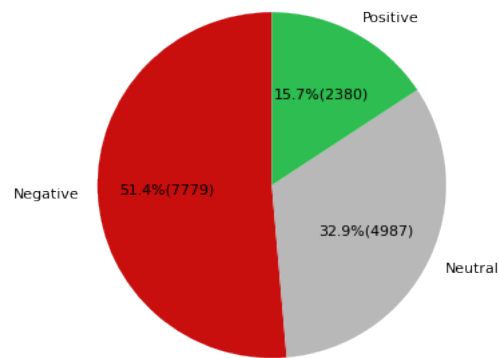
With the help of our sentiment analysis algorithm, every comment below every topic will be analyzed through the algorithm and results graphs will be created and will be shown in this admin page.

# Istanbul Medipol University

## School of Engineering and Natural Sciences

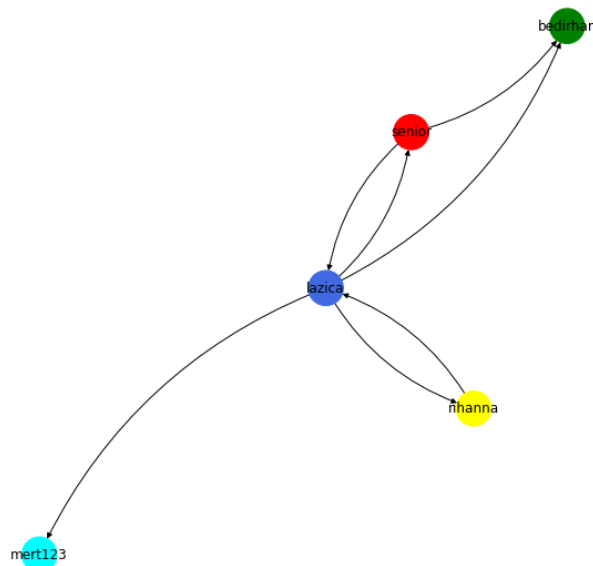
### Graduation Project

Distribution of Sentiment



*Image 9. Sentiment Result*

After sentiment analysis algorithm first step was to create friendship matrix. According to this matrix, graphs will be created and will be shared in admin page as seen below.



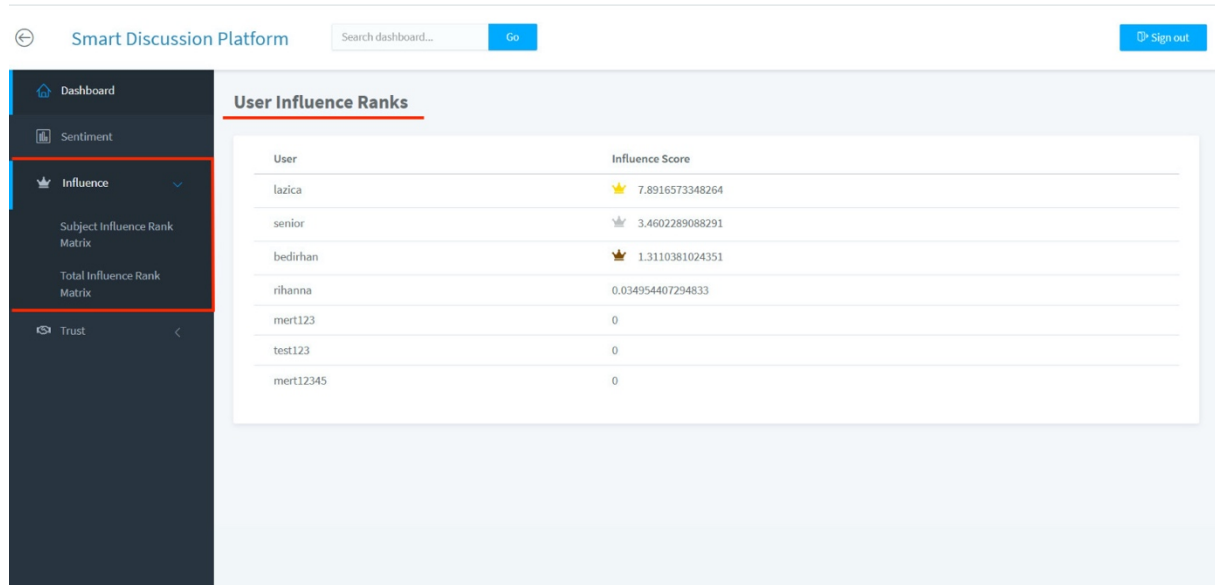
*Image 10. Follower-User Graph*

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

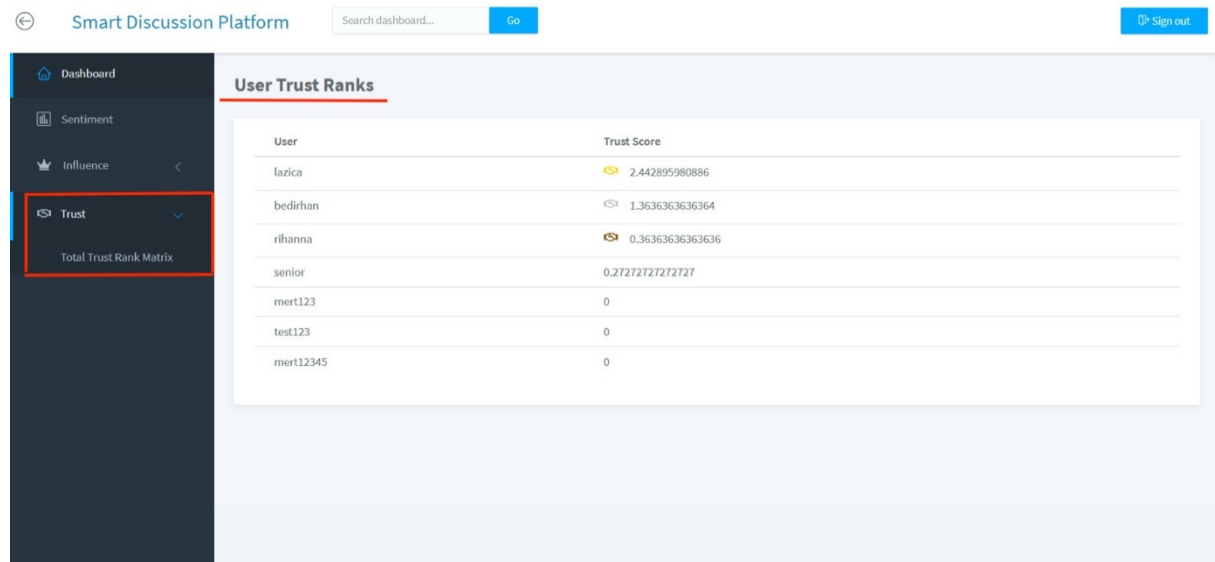
After that, influence ranks are ordered according to result of Influence scores. Below is an example from admin page which shows the rank of the users according to Influence.



| User      | Influence Score   |
|-----------|-------------------|
| lazica    | 7.8916573348264   |
| senior    | 3.460228908291    |
| bedirhan  | 1.3110381024351   |
| rihanna   | 0.034954407294833 |
| mert123   | 0                 |
| test123   | 0                 |
| mert12345 | 0                 |

Image 11. Influence Rank

Trust ranks are ordered according to result of influence scores. Below is an example from admin page which shows the rank of the users according to Trust.



| User      | Trust Score      |
|-----------|------------------|
| lazica    | 2.442895980886   |
| bedirhan  | 1.3636363636364  |
| rihanna   | 0.36363636363636 |
| senior    | 0.27272727272727 |
| mert123   | 0                |
| test123   | 0                |
| mert12345 | 0                |

Image 12. Trust Rank



# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

Lastly we have created a user evaluation system for users to evaluate website. As seen below:

Feedback

Your Comment:

Rate Us

Excellent

Send

Close

*Image 13. User Evaluation Form*

## 9. DISCUSSION

In the project, mainly three algorithm development is aimed along with the website creation. Fundamental of all these algorithms were first collecting data, which can be taken from our own website or any social media and creating a good preprocessing method for sentiment analysis. So according to these aims and fundamentals, data collection and preprocessing implementation were done through the methods explained in the experiment part. This step was important also since preprocessing steps are crucial because it transforms text into a more digestible and understandable form so that our algorithms can perform better.

After that, it is aimed to build the sentiment analysis algorithm. This algorithm also serves a main functionality in our project and website. It helps to analyze the comments and expressions under the topics on our website. It is seen from the results from all methods that are done for validation, approximately 80% accuracy is reached which is a satisfactory result in case of our aims. Higher accuracy can be reached by implementing different preprocessing methods and with different training methods. This algorithm helped us to evaluate the comments under every topic to see the user reactions to a topic.

After that, website platform is done. Process of website construction divided into two part: Front-end and Back-end. Difficulties are mainly faced in the front-end part. A smooth and compatible environment is aimed for users. Also, for the website, a user evaluation system is done to evaluate the satisfactory of the users from website. %90 percent is achieved from this evaluation.

# Istanbul Medipol University

## School of Engineering and Natural Sciences

### Graduation Project

After website creation, last algorithms Trust and Influence are implemented. These algorithms are implemented after website creation since the data that is needed for algorithms was taken from website. Through the process of Trust and Influence algorithms it is observed that there are lots of criteria and implementation for this standard which is differ from platform to platform or according to aims. Because of that, lots of source and papers were analyzed and a unique algorithm were created and integrated for our own platform. Lastly, these interaction and algorithm results are presented in a different page created for this aim.

#### 10.RISKS AND B PLAN

| Risk   | B Plan   |
|--|--|
| Insufficient labeled data for sentiment analysis | Merging multiple labeled data                                |
| Getting low accuracy from an algorithm           | Alternative algorithm or hybrid algorithm can be implemented |
| Scraper Error                                    | Alternative Scraper or Twitter API can be used               |
| Domain server collapsing                         | New domain server  |

#### 11.CONCLUSION

On the way of our project we have created a “Smart Discussion Platform” that allows people to create their own social network and create their own topics and discuss them. On the other hand, we created algorithms first to analyze these topics in terms of positivity and negativity and comments under these topics with our sentiment analysis algorithms. Then, we analyze the interaction between people with our two algorithm “Influence” and “Trust”.

#### 12.REFERENCES

- [1] Chawan, Pramila. (2012). Sentiment Analysis and Influence Tracking using Twitter. International Journal of Advanced Research in Computer Science and Electronics Engineering. 1.
- [2] J. Cossu, N. Dugué and V. Labatut, "Detecting Real-World Influence through Twitter," *2015 Second European Network Intelligence Conference*, Karlskrona, 2015, pp. 83-90.
- [3] Sanandres, Eliana & Llanos, Raimundo & Madariaga Orozco, Camilo. (2018). Topic Modeling of Twitter Conversations.
- [4] Doan, S., Yang, E.W., Tilak, S.S. *et al.* Extracting health-related causality from twitter messages using natural language processing. *BMC Med Inform Decis Mak* 19, 79 (2019).
- [5] Alowisheq, Areeb & Alrajebah, Nora & Alrumikhani, Asma & Al-Shamrani, Ghadeer



# **Istanbul Medipol University**

## **School of Engineering and Natural Sciences**

### **Graduation Project**

& Shaabi, Maha & Al-Nufaisi, Muneera & Alnasser, Ahad & Al-Humoud, Sarah. (2017). Investigating the Relationship Between Trust and Sentiment Agreement in Arab Twitter Users. 236-245.

[6] K. Tago and Q. Jin, "Influence analysis of emotional behaviors and user relationships based on Twitter data," in *Tsinghua Science and Technology*, vol. 23, no. 1, pp. 104-113, Feb. 2018, doi: 10.26599/TST.2018.9010012.

[7] Li, P., Zhao, W., Yang, J., & Wu, J. (2019, August). CoTrRank: Trust Evaluation of Users and Tweets. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}.

[8] Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12), 2521–2535.

[9] Cano Basave, A. E.; Mazumdar, S. and Ciravegna, F. (2014). Social influence analysis in microblogging platforms – a topic-sensitive based approach. *Semantic Web*, 5(5) pp. 357–403.

[10] Sung, J., Moon, S., & Lee, J.-G. (2013). The Influence in Twitter: Are They Really Influenced? In *Behavior and Social Computing* (pp. 95–105). Springer International Publishing.

[11] Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics - SOMA '10. the First Workshop.

[12] Grant, Christan & George, Clint & Jenneisch, Chris & Wilson, Joseph. (2011). Online Topic Modeling for Real-Time Twitter Search.

[13] Jonsson, Elias. "An Evaluation of Topic Modelling Techniques for Twitter." (2016).

[14] Ruan, Y., Durrese, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, 145, 207–218

[15] Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., & Tesconi, M. (2017, October). Cross-Media Learning for Image Sentiment Analysis in the Wild. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). 2017 IEEE International Conference on Computer Vision Workshop (ICCVW).

[16] Go, Alec & Bhayani, Richa & Huang, Lei. (2009). Twitter sentiment classification using distant supervision. *Processing*. 150.



# **Istanbul Medipol University**

## **School of Engineering and Natural Sciences**

### **Graduation Project**

[17] Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank. Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10. the third ACM international conference.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[19] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27(1):415–444, 2001.



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**  
**PROJECT ACTIVITIES AND WORK PLAN**

| Work and Activity                       | Responsible Group Member           | Timeline |         |         |         |         |         |         |         |         |          |          |          |          |          |
|---|------------------------------------|----------|---------|---------|---------|---------|---------|---------|---------|---------|----------|----------|----------|----------|----------|
|   |                                    | 1. week  | 2. week | 3. week | 4. week | 5. week | 6. week | 7. week | 8. week | 9. week | 10. week | 11. week | 12. week | 13. week | 14. week |
| 1. Literature Review                    | Bedirhan Gergin<br>Mert Can Çakmak | ✓        | ✓       | ✓       | ✓       | ✓       | ✓       |         |         |         |          |          |          |          |          |
| 2. Developing Back-End of the Website   | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         | ✓       | ✓       | ✓       | ✓       |          |          |          |          |          |
| 3. Developing Front-End of the Website  | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         |         | ✓       | ✓       | ✓       | ✓        |          |          |          |          |
| 4. Sentiment Analysis Algorithm Design  | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         | ✓       | ✓       | ✓       | ✓       |          |          |          |          |          |
| 5. Trust and Influence Algorithm Design | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         |         |         | ✓       | ✓       | ✓        | ✓        | ✓        |          |          |
| 6. Invitation Policy Algorithm Design   | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         |         |         | ✓       | ✓       | ✓        | ✓        | ✓        |          |          |
| 7. Implementation of the Algorithms     | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         |         |         |         |         |          | ✓        | ✓        | ✓        |          |
| 8. Testing of the Algorithms            | Bedirhan Gergin<br>Mert Can Çakmak |          |         |         |         |         |         |         |         |         |          |          | ✓        | ✓        | ✓        |



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

|   |                                    |  |  |  |  |  |  |  |  |  |  |  |   |   |   |
|---|------------------------------------|--|--|--|--|--|--|--|--|--|--|--|---|---|---|
| <b>9. System Integration and Testing</b>        | Bedirhan Gergin<br>Mert Can Çakmak |  |  |  |  |  |  |  |  |  |  |  | ✓ | ✓ | ✓ |
| <b>10. Releasing and Testing of the Website</b> | Bedirhan Gergin<br>Mert Can Çakmak |  |  |  |  |  |  |  |  |  |  |  |   | ✓ | ✓ |



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

**LIST OF WORK PACKAGES**

| WP No | Detailed Definition of Work and Activity   |
|-------|--|
| 1     | Existing works and papers about Data Mining, Sentiment Analysis, Trust and Influence will be examined.   |
| 2     | Database design and server related parts will be done.   |
| 3     | User interface design will be done.  |
| 4     | Algorithm for the sentiment analysis will be created.  |
| 5     | Algorithm for the trust and influence analysis will be created.  |
| 6     | Algorithm that enables to invite people from social media and microblogs will be created. This will include a hybrid system that include both sentiment and influence analysis of other platforms. |
| 7     | Algorithms will be implemented with sample data.   |
| 8     | Correctness of the algorithms will be evaluated by comparing the results with existing techniques.   |
| 9     | Algorithms will be combined with platform after integrating them into to the system.   |
| 10    | Platform will be tested as a whole. After that, it will be released to live users.   |



**Istanbul Medipol University**  
**School of Engineering and Natural Sciences**  
**Graduation Project**

| Work package | Target   | Measurable outcome               | Contribution to overall success(%) |
|--------------|--|----------------------------------|------------------------------------|
| 1            | Gathering information for the project implementation and algorithms. | Algorithms                       | %10                                |
| 2            | Designing server side of the platform.                               | Database and Server Organization | %10                                |
| 3            | Designing user side of the platform.                                 | Interface of Website             | %10                                |
| 4            | Creating sentiment analysis algorithm.                               | Accuracy of Algorithm            | %15                                |
| 5            | Creating influence and trust analysis algorithm.                     | Accuracy of Algorithm            | %15                                |
| 6            | Creating algorithm for the invitation policy.                        | Accuracy of Algorithm            | %15                                |
| 7            | Applying created algorithms.   | Accuracy of Algorithms           | %5                                 |
| 8            | Evaluating the correctness of the algorithms.                        | Accuracy of Algorithms           | %5                                 |
| 9            | Creating the whole system with integrating algorithms.               | Accuracy of Algorithm in Website | %5                                 |
| 10           | Presenting website to users and testing them with real users.        | Accuracy and User Experience     | %10                                |

| WORK PACKAGE DISTRIBUTION |     |     |     |     |     |     |
|---------------------------|-----|-----|-----|-----|-----|-----|
| Project Member            | WP1 | WP2 | WP3 | WP4 | WP5 | WP6 |
| Mert Can ÇAKMAK           | %50 | %50 | %50 | %50 | %50 | %50 |
| Bedirhan GERİN            | %50 | %50 | %50 | %50 | %50 | %50 |
| Total                     | 100 | 100 | 100 | 100 | 100 | 100 |