



T.C

ALTINBAS UNIVERSITY

Graduate School of Science and Engineering

Information Technologies

**FEATURE SELECTION USING RANKING
ALGORITHMS**

Afaf Abdulhamed Omar Elraheibi

M.Sc. Thesis

Supervised by Prof. Dr. Oğuz Bayat

Istanbul,2019

FEATURE SELECTION USING RANKING ALGORITHMS

By

Afaf Abdulhamed Omar Elraheibi

Information Technologies

Submitted to the Graduate Faculty of
Altinbas Universities in partial fulfillment
of the requirements for the degree of
Master of Computer Engineering.

ALTINBAS UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of

Academic Title Name SURNAME

Co-Supervisor

Academic Title Name SURNAME

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Academic Title Name SURNAME Faculty, University _____

Academic Title Name SURNAME Faculty, University _____

Academic Title Name SURNAME Faculty, University _____

Academic Title Name SURNAME Faculty, University _____

Academic Title Name SURNAME Faculty, University _____

I certify that this thesis satisfies all the requirements as a thesis for the degree of

Academic Title Name SURNAME

Head of Department

Approval Date of Graduate School of
Science and Engineering: ____/____/____

Academic Title Name SURNAME

Director

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

Afaf Abdulhamed Omar Elraheibi

ACKNOWLEDGEMENTS

I'm deeply grateful to my supervisor Prof. Dr. Oğuz Bayat for his guidance, his time, and invaluable comments on the thesis.

I would like to express my deepest appreciation to my friends for their unconditional support.

Also I would thank my country Libya for the unlimited support.

My final words go to my family. I want to thank my family, whose love and guidance is with me in whatever I pursue.

ÖZET

ÖZELLİL SEÇİMİ SIRALAMA ALGORİTMALARINI KULLANMA

Afaf Abdulhamed Omar Elraheibi

Yüksek Lisans Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi

Danışman: Prof. Dr. Oğuz Bayat

Tarih: Ağustos, 2019

Sayfalar: 57

Günümüzde veriler birçok biçimde ve inanılmaz boyutta saklanır, bu kadar veriyi analiz etmek için ve Verilerin makine tarafından anlaşılabilir kılınması için yeni yöntemler ve algoritmalar geliştirilir, Orijinal verileri “Özellikler” biçiminde gösteren daha basit bir forma dönüştürülmelidir. Verilerin özelliklere dönüştürüldüğü sürece “Özellik Çıkarma” işlemi denir. Ancak bir çok özellik var ve problem için en verimli olanı kullanmak gerekiyor, bu yüzden özellik seçim süreci önemli. Kaur ve diğ. [7] araştırmalarında, özellikleri sıralamak ve çalışmalarına en uygun özellik grubunu seçmek için Çok Kriterli Karar Destek Yöntemleri “MCDM” den biri olan AHP-topsis algoritmasını kullandılar, Bu araştırmada çalışmalarına daha fazla MCDM algoritması uygulanarak devam edilecek, özellikleri sıralamak için sonra da bu sonuçları karşılaştırmak için karşılaştırıyoruz. Kullanılacak en uyumlu özellikler seti

Anahtar kelimeler: Özellik çıkarma, Özellik seçimi, MCDM, ANP, VIKOR, ELECTRE, PROMETHE

ABSTRACT

FEATURE SELECTION USING RANKING ALGORITHMS

Afaf Abdulhamed Omar Elraheibi

M.Sc. Information Technologies, Altinbas University,

Supervisor: Prof. Dr. Oğuz Bayat

Date: August, 2019

Pages: 57

Nowadays the data is stored in many forms and with incredible size, new methods and algorithms are developed to analyze such amount of data, to make the data understandable by the machine, it should be converted to a simpler form that represents the original data in the shape of “Features”. The process in which the data is converted to features is called the “Feature Extraction” process. But there are a lot of features and just the most efficient to the problem should be used, therefore the features selection process is important. Kaur et al. in their research [7] used the AHP-topsis algorithm which is one of the Multi Criteria Decision Support Methods “MCDM” to rank the features and select the most appropriate features group to their work, in this research their work will be continued by applying more MCDM algorithms, to rank the features then we compare this results to extract The most compatible set of features to be used.

Keywords: Feature extraction, Feature selection, MCDM, ANP, VIKOR, ELECTRE, PROMETHE

TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACTC	vi
LIST OF TABLES	x
LIST OF FIGURES	xii
1. INTRODUCTION	1
1.1 LITERATURE REVIEW	2
1.2 THE PROBLEM DEFINITION	8
1.3 WHAT ARE THE FEATURES?	9
1.4 WHAT ARE THE MEASURES?	10
2. THEORITICAL FRAMEWORK	12
2.1 WHAT IS MCDM?	12
2.2 METHODS	13
2.2.1 ANP Method.....	13
2.2.2 ELECTRE Method	17
2.2.3 VIKOR Method	20
2.2.4 PROMETHEE Method.....	22
3. METHODOLOG	28
3.1 RESEARCH GOAL	28
3.1.1 Alternatives.....	28

3.1.2	Criteria	28
3.2	PROBLEM SOLVING.....	29
3.2.1	Solution of Problem With ANP Method	29
3.2.2	Solving The Problem With ELECTRE Method	31
3.2.3	Solution Of Problem With VIKOR Method.....	38
3.2.4	Solution Of Problem With PROMETHEE Method	40
3.3	COMPARISON.....	42
4.	CONCLUSION.....	44
	REFERENCES.....	45

LIST OF TABLES

	<u>Pages</u>
Table 1.1: Ranking using χ^2 (Chi-Squared) method [8].....	3
Table 1.2: The top ten features [9].....	4
Table 1.3: The accuracy matrix of each features groups [7].....	8
Table 2.1: The evaluation table [23].....	24
Table 2.2: Preference functions criteria [22]	26
Table 3.1: Accuracy table [7].....	29
Table 3.2: ANP ranking result	31
Table 3.3: Normalized original matrix.....	31
Table 3.4: Weighted normalized matrix	32
Table 3.5: C Matrix.....	35
Table 3.6: D Matrix.....	36
Table 3.7: E Matrix	36
Table 3.8: F Matrix	37
Table 3.9: Ck and Dk values.....	37
Table 3.10: ELECTRE Ranking Result	38
Table 3.11: F+I and F- I' values	38
Table 3.12: Sj and Rj values	39
Table 3.13: Qj ' values	39

Table 3.14: VIKOR Ranking Result..... 39

Table 3.15: PROMETHEE Ranking Result 39

Table 3.16: Methods Ranking Results 42

Table 3.17: Feature Ranking After Filtering..... 43



LIST OF FIGURES

	<u>Pages</u>
Figure 1.1: Text Mining Model [1].....	1
Figure 1.2: “GRA” Theory and Multi-Objective Optimization Method [10].....	5
Figure 1.3: Peng et. Al, Suggested Schema [9]	6
Figure 1.4: Singh et Al. Methodology [12]	7
Figure 1.5: Features Hierarchy [7].....	9
Figure 2.1: The Structure of ANP [16]	14
Figure 2.2: Difference In Structure Between AHP (left) and ANP (Right). [17].....	16
Figure 3.1: Network Structure Defined Among The Criteria And Alternatives.....	30
Figure 3.2: Super Decision Screen Display Of Priorities Of Alternatives	30
Figure 3.3: Visual PROMETHEE Academic" Program Interface.....	40

1. INTRODUCTION

As a result of the large amount of text data produced by the Social Media platforms, an urgent need for increasing the research to extract the knowledge from these data appeared; with high precision of the artificial intelligence algorithms, Which have become widely used, and lead to appear many applications in this field such: as emotion analysis, text classification, text synthesis, plagiarism...

All of these applications have been followed through a common approach, with sequential steps as figure (1.1) shows: [1]

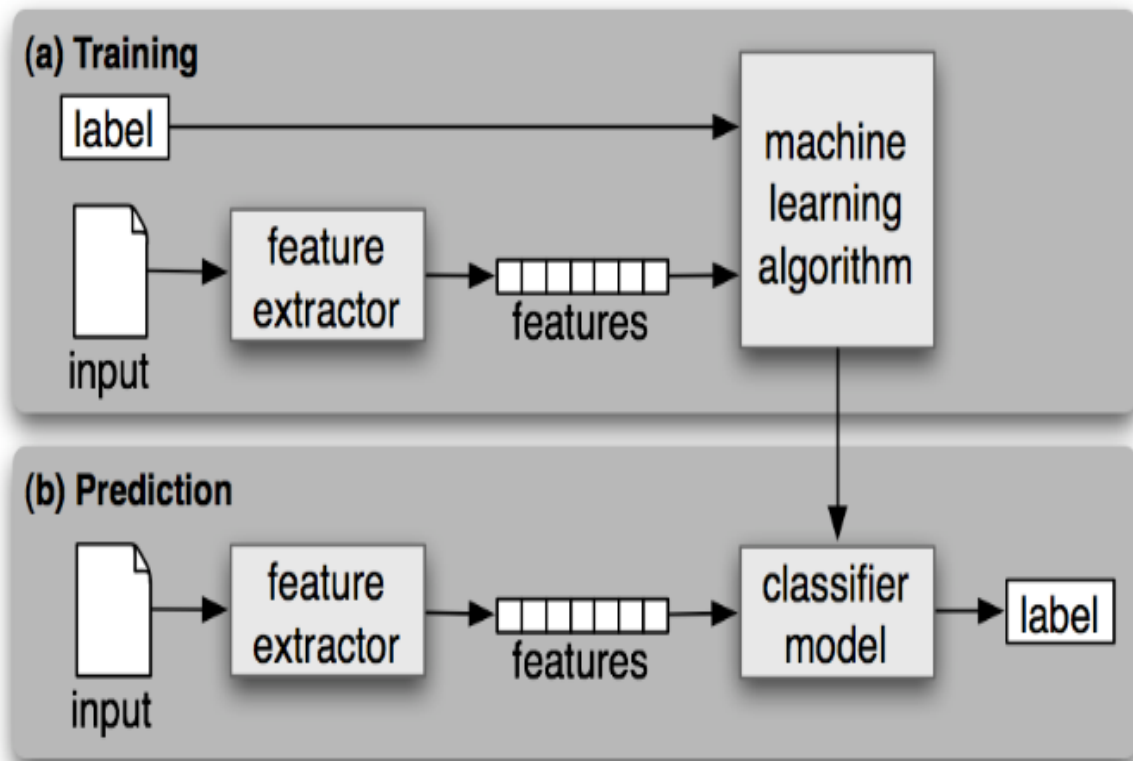


Figure 1.1: Text mining model [1]

This research will held in the feature extraction step, which the process that reduces dimensions [2]. When an algorithm's income is too large to be easily manipulated, and is expected to constitute a surplus in data, which rise the computing, processing costs, and the use of computer memory

without a corresponding return to that cost. The data is then converted to a simpler form that represents the original data in the shape of “Features”. The process in which the data is converted to features is called the “Feature Extraction” process.

Inside this step there a sub process called “Feature selection” , which is a technique used extensively in machine learning to select a subset of features for a data set in order to build a stable learning model [3]. The features selection process used when we have a lot of features and we just want to use the most efficient to our problem. According to Bermingham et al. “The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information” [5]; For example if there is a tweet, and his author want to be known to detect the “plagiarism”, just the features related with the author style from previous tweets is selecting, in order to teach the machine to be able to predict the real tweet author. this research will focus on the “Feature selection” in the Authorship Attribution problem, which is a famous problem in the text mining appeared in 1887 when Mendenhall first invented the idea of Counting Features to indicate the personality of the author [4]

There are many methods to select the features, Guyon et al. suggested “The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate” [6], that if we are working in just one measure “Error rate”, but what if we are working in many measures: Precision, Recall, F -score, False Rejection Rate (FRR), False Acceptance Rate (FAR), and Co-efficient of Variance (CV) like in Kaur et al. article [7] who used the AHP-topsis algorithm which is one of the Multi Criteria Decision Support Methods “MCDM” to rank the features and select the most appropriate features group to their work, this research will continue their work by applying more MCDM algorithms, to rank the features then comparing the results to extract The most compatible set of features to be used.

1.1 LITERATURE REVIEW

Benevenuto et al. in their research to detect spammers in twitter, used information gain and χ^2 (Chi-Squared) to select the most effective features , the two methods are exists in WEKA tool, the ranking for the features by using the two methods are nearly, so the authors omitted the information gain method and used just the χ^2 (Chi-Squared) method. In the table (1.1) the top ten ranking

features according to the used method [8] the authors used 62 features so they divided the features to groups each one contains 10 features according to the χ^2 (Chi-Squared) method ranking, then they calculated the classifier according to each group [8].

Table 1.1: Ranking using χ^2 (Chi-Squared) method [8]

Position	χ^2 ranking
1	Fraction of tweets with URLs
2	Age of the user account
3	Average number of URLs per tweet
4	Fraction of followers per followees
5	Fraction of tweets the user had replied
6	Number of tweets the user replied
7	Number of tweets the user replied a reply
8	Number of followees
9	Number of followers
10	Average number of hashtags per tweet

Criado et al. in their research about authorship attribution on Facebook dataset, Criado et al. used five feature sets: Structural, POS, Semantic, Category, Style feature sets, to make the features selection the information gain (IG) technique have been used, the information gain technique are used frequently in machine learning algorithms especially the decision tree algorithm; and the feature which got the higher IG are considered than the other features; Criado et al. calculated the IG for all the features and they trained the classification algorithms SVM and J48 by using the top 250 features which got the highest IG from 650 features (the total of all the features from the five sets), the table (1.2) shows the top ten features [9].

Table 1.2: The top ten features [9]

Feature Type	Feature description	IG
Structural	Av. Influence	0.68
Structural	Messages	0.48
Semantic	General and Abstract Terms	0.37
Semantic	Measurement	0.36
Semantic	Social Actions, states and processes	0.35
Semantic	Money generally	0.34
POS	Base from of lexical verb (e.g., give)	0.34
Semantic	Degree (i.e., intensifier terms)	0.34
Semantic	Quantities	0.33
Style	Av. Typing	0.32

Ma et. Al, in their research focused in selecting the optimal attributes for the decision making problems, where these attributes are objective and subjective, to do that attributes analysis framework from two step screening procedure using Grey Relational Analysis “GRA” theory and multi-objective optimization method suggested, as figure (1.2) shows [10]:

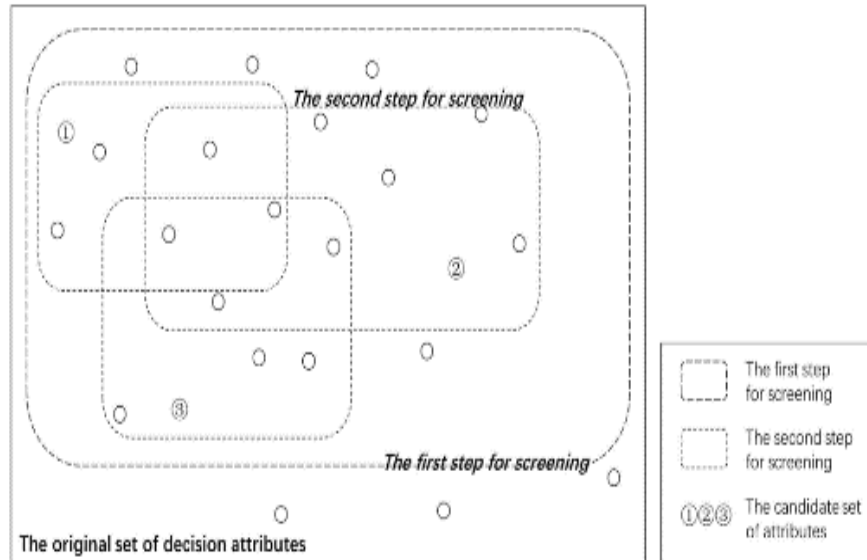


Figure 1.2: “GRA” Theory and Multi-Objective Optimization Method [10]

Peng et. Al, suggested many steps features selection schema as figure (1.3) shows [9]; where many datasets from many fields and sizes selected; second applying 10 fold cross validation on the datasets, then three techniques used to select the feature: WrapperSubsetEval, CfsSubsetEval, and ConsistencySubsetEval, which engaged in WEKA tool; third step apply MCDM methods to evaluate the previous three techniques and choose the best one from them; fourth using the selected features in classification; finally the classification result compared with the traditional model to examine the prediction accuracy improvement [11]. The proposed schema used the rank features not feature groups, also the MCDM methods here used to rank the techniques not the features.

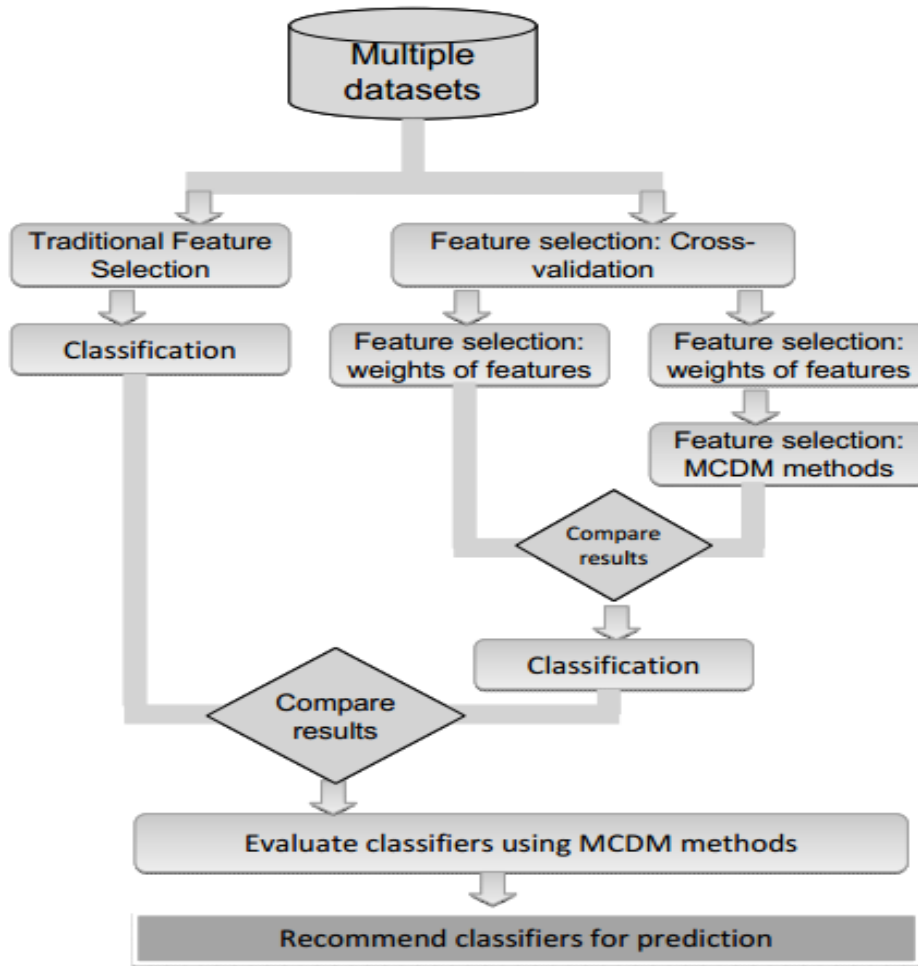


Figure 1.3: Peng et. Al, Suggested Schema [9]

Singh et al. in their research minimize the number of extracted features from the processed packets, which consume process and time, used one of the MCDM methods called “TOPSIS” by MATLAB tool, to rank ten features selection techniques used to analyze the famous dataset “KDD network” [12]. The used research methodology shown in figure (1.4):

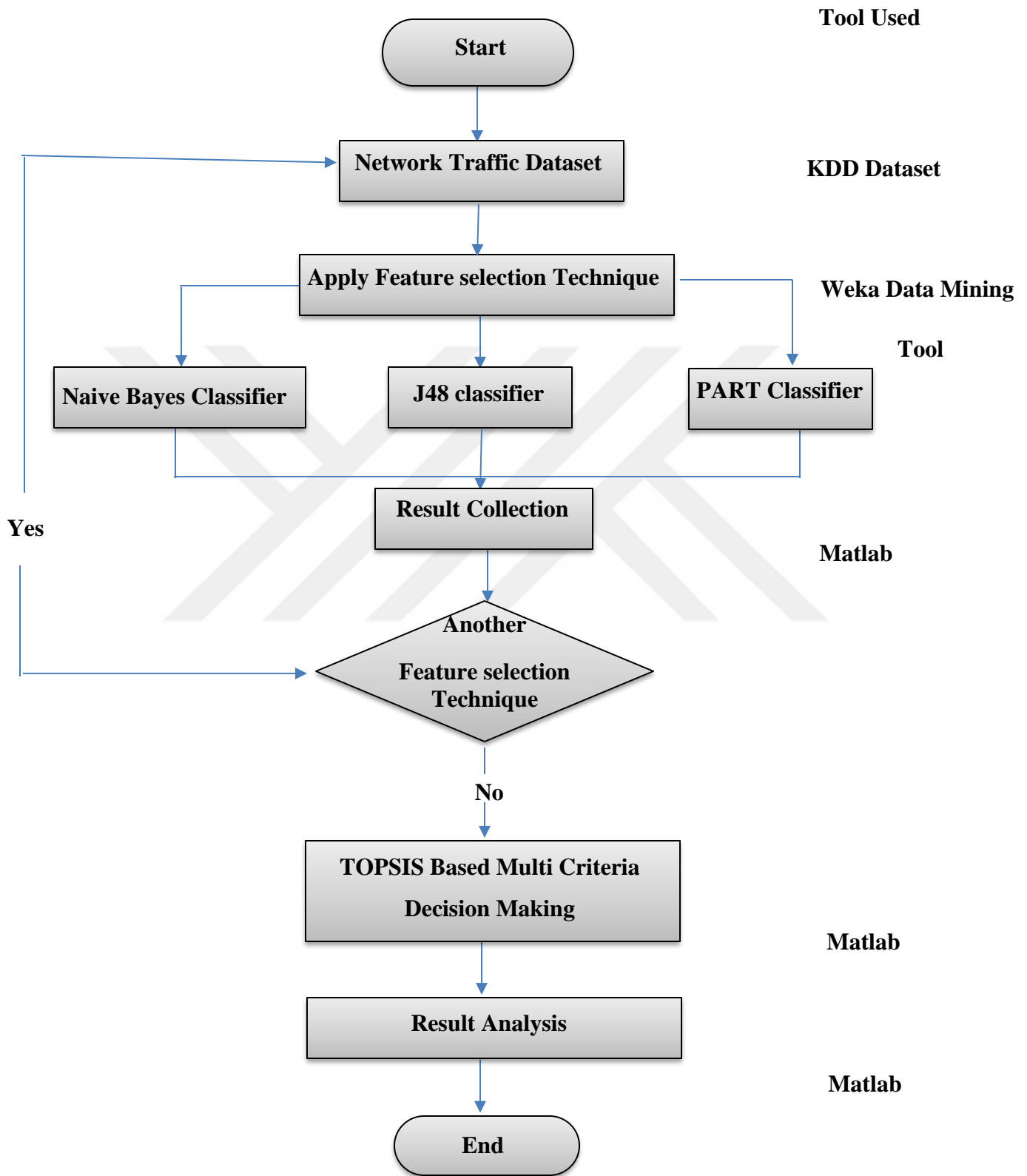


Figure 1.4: Singh et al. Methodology [12]

Kaur et al. in their research tried to verify the author of a specific tweet, by extracting feature groups from the tweets, and use these features to learn the machine, to detect the new tweet' author, Kaur et al. used many group features as alternatives and ranking them using one of the MCDM methods called "AHP-TOPSIS" according to many measures or criteria [7].

In this research we will continue Kaur et al. works by using other MCDM methods to rank the features groups that will help us to compare the ranking result to get best accuracy.

1.2 THE PROBLEM DEFINITION

Kaur et al. in their research tried to verify the author of a specific tweet, by extracting feature groups from the tweets and learn the machine on these features, to detect the new tweet' author, Kaur et al. used many measures to detect their prediction accuracy; table (1.3) shows the accuracy matrix of each features groups [7]:

Table 1.3: The Accuracy Matrix of Each Features Groups [7]

	TRUE RATE ACCURACY			FALSE RATE ACCURACY		Co-efficient of Variance
	<i>F</i>	<i>R</i>	<i>P</i>	<i>FAR</i>	<i>FRR</i>	CV
C	87.50	84.00	91.30	7.41	16.00	88.46
U	90.57	96.00	85.71	14.81	4.00	90.38
B	86.27	84.61	88.00	10.71	15.38	87.04
L1	79.36	100.00	65.78	48.14	0.00	75.00
T	83.02	88.00	78.57	22.22	12.00	82.69
S	77.55	76.02	79.16	18.52	24.02	78.85
N	84.74	100.0	73.53	33.33	0.00	82.69

The items in the lines are the features groups : Char n-gram "c", Unigram "U", Bigram "B", SPATIUM-L1 "L1", tfidf "T", Stylometric "S", NMF "N"; which represent the alternatives that we want to rank by using the MCDM methods

The items in the columns: F –score “F”, Recall “R”, Precision “P”, False Acceptance Rate “FAR”, False Rejection Rate “FRR”, Accuracy “A” which represent the criteria that according the alternatives will select.

The ranking of these alternatives will done through six criteria, the above table shows the evaluation degree for each feature group according to each criteria; the table obtained from a previous study (Kaur et al. 2018). This table will be used in each MCDM method to rank the alternatives. Ranking these features help us to select the features which can achieve the highest accuracy, and the lowest error rate.

1.3 WHAT ARE THE FEATURES?

Kaur et al. classified the features that can be extracted from the text as figure (1.5) shows [7]:

And Using the following features [7]:

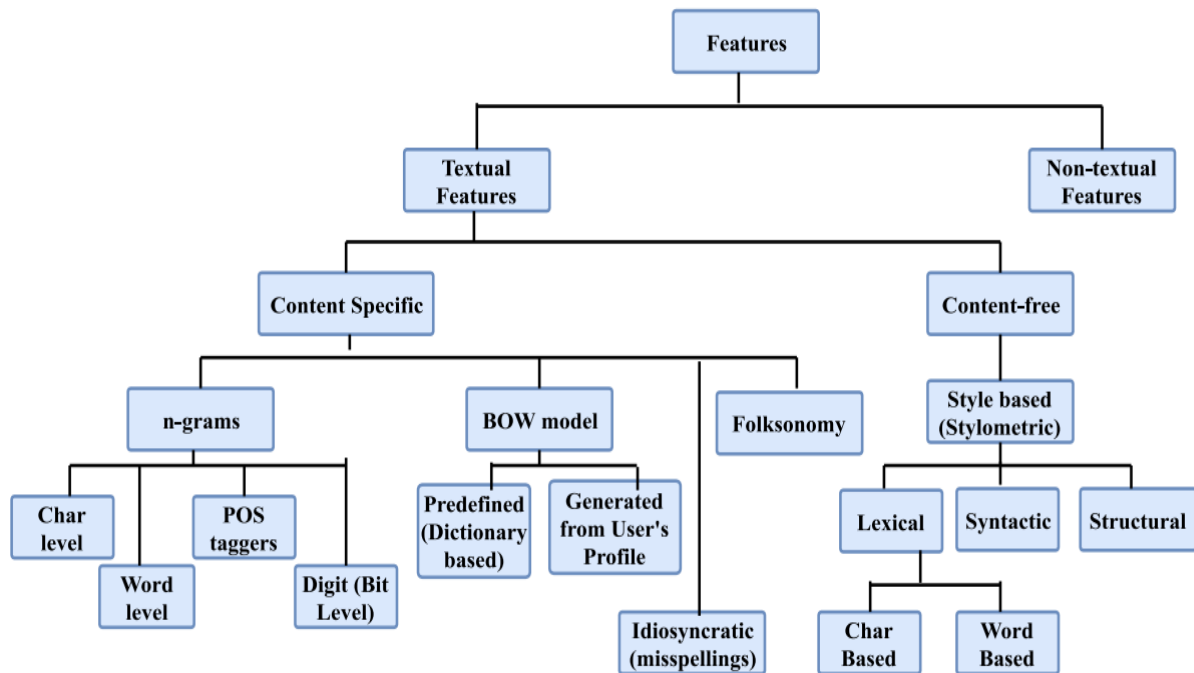


Figure 1.5: Features Hierarchy [7]

- Char n-grams: considering each n-gram character as a unit to process them, for example the sentence “we are” :

The char 1-gram will be ‘w’,’e’,’a’,’r’,’e’

The char 2- gram will be ‘we’,’e_’,’_a’,’ar’,’re’

- Unigram: equal to char 1-gram
- Bigram: equal to char 2-gram
- SPATIUM L1: according to Kaur et al. “Efficient unsupervised authorship verification technique was proposed by Kocher and Savoy (2017) using a simple distance comparison measure called SPATIUM-L1 for the extracted k-most frequent words as features. The proposed technique used statistical analysis and did not involve a learning step to train and define the values of parameters. Threshold values were varied to compute the values of different performance parameters.” [7]
- TF-IDF: is an acronym for “term frequency/inverse document frequency “ the term frequency mean how many a term accrued in the text, and the inverse document frequency represent the number of texts that the term is in .
- Stylometric features : are content free features, contain many features group:
 - Lexical features: a set of items related with word in the text
 - Syntactic features: a set of punctuations and function words which are important because its describes how the words grammatically related together in the sentence
 - Structural features: the format and organization of a text
- NMF: is an acronym for Non-Negative Matrix Factorization, which used in topic modeling

1.4 WHAT ARE THE MEASURES?

The author used the following measure to detect the classification accuracy [7]:

- Precision: the percentage of “True Positive” occurrences on the sum of the “True Positive”+” False Positive” occurrences in eq (1.1)

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (1.1)$$

- Recall: the percentage of “True Positive” occurrences on the sum of the “True Positive”+” False Negative” occurrences in eq (1.2)

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \quad (1.2)$$

- F –score: it’s a measure depend on the “Precision” and “Recall” measures occurrences in eq (1.3)

$$\text{F-measure} = 2 * ((\text{precision} * \text{Recall}) / (\text{precision} + \text{Recall})) \quad (1.3)$$

- False Rejection Rate (FRR) : the percentage of “False Negative” occurrences on the sum of the “True Positive”+” False Negative” occurrences in eq (1.4)

$$\text{FRR} = \text{FN}/(\text{FN} + \text{TP}) \quad (1.4)$$

- False Acceptance Rate (FAR): the percentage of “False Positive” occurrences on the sum of the “True Negative”+” False Positive” occurrences in eq (1.5)

$$\text{FAR} = \text{FP}/(\text{FP} + \text{TN}) \quad (1.5)$$

- Co-efficient of Variance (CV): “computed using the standard deviation and mean values of respective performance parameter of all users as follows” in eq (1.6)

$$\text{CV} = \sigma/\mu \quad (1.6)$$

2. THEORITICAL FRAMEWORK

2.1 WHAT IS MCDM?

Jahan and Edwards define it as following “MCDA is an acronym that stands for Multiple Criteria Decision Analysis/Aiding, and it is sometimes referred as MCDM (Multiple Criteria Decision Making). It is a sub discipline of operational research. Operational research is often considered to be a subfield of mathematics that applies advanced analytical methods to get optimal or near-optimal solutions in complex decision-making problems. Operational research is focused on practical problems in marketing, manufacturing, transportation, information technology (IT) and other fields. Therefore operational research overlaps with other disciplines, particularly operations management and engineering science. MCDA is a sub discipline of operational research that explicitly deals with decision problems that use multiple criteria to determine the best possible solution” [13]

The MCDM include the following steps: [14]

- Identify alternatives:

The alternatives are the group of suggestions that the decisions will built on them, and represent all the possible solutions. Where formatting the alternatives group done due to the decision goal, for example in building a university the alternatives are the locations. And the alternatives must be clear and each one represent an entity by itself. The alternatives in this research are the “features” which will ranked.

- Define criteria:

The criteria represent all the viewpoints that effect on the suggested decisions; which represent the needs and the goals which should be in the alternative; defining the criteria is the process of collecting the enough and necessary information about the expected performance for the alternative, the criteria should be formulated by quantitative or qualitative mathematical forms; and should not be incomplete or repeated in many names under the same meanings. The criteria in this research are the “measures” which used to rank the features.

- Define the criteria’s weights:

Each criteria has a different importance and effect on the decision making, for that a weight should be given to each criteria to represent its importance, the weight can be a percentage or

number. In fact giving the weight on of the complex challenge in the MCDM, because of the self-preferences and the self-impact of the evaluator.

- Selection the method to evaluate alternatives:

There are many methods used in the MCDM: (AHP, ANP, TOPSIS, ELECTRE, PROMETHEE and VIKOR)

- Evaluate alternatives against criteria:

The MCDM methods help to rank the alternatives and give the decision makers a viewpoint about the appropriate of each alternative according criteria

- Validate solutions against problem statement

2.2 METHODS

As mentioned above MCDM method using to get the optimize solution for a decision problems with multiple criteria, actually MCDM a ranking approach, where it helps to rank group of alternatives based on multiple criteria values due to the most ability of this alternative to do the task; the criteria values (or weight) put by the experts in the field where not all of the criteria have the same importance.

In following the fourth most important MCDM methods (ANP, ELECTRE, PROMETHEE and VIKOR) will be explained to understand the mechanism of each method and how can we use it to rank the alternative to solve our problem later.

2.2.1 ANP Method

The Analytic Network Process a more generalized model of the Analytic Network Process (AHP). Invented by Thomas L.Saaty in 1996. [15]

The method of network analysis is one of the multivariate analysis that uses the structure of network to model the problem and the pairwise comparisons to make the relationships in the structure.

The ANP ranks the alternatives group which have number of criteria. There are a preferences established between the criteria and alternatives done by the pairwise comparisons. The alternative which ranked as the best by this method is the most suitable one for the DM.

The structure of ANP consists of clusters and nodes, each cluster contains many nodes which connected together in the both directions as seen in the figure (2.1); each cluster includes one of the problem elements: problem goals, alternatives, and criteria. Where grouping the nodes in a cluster is one of the differences between ANP and AHP. [16] Helps not just to compare priorities between nodes but also between clusters.

The ANP network are representing in a Matrix contains all the nodes vertically and horizontally and each non-zero element of the matrix represent the weight and connection from a node horizontally to other node inside the network vertically, this matrix after preparing are called super matrix, which contains all the related important for node to other nodes or cluster other clusters.

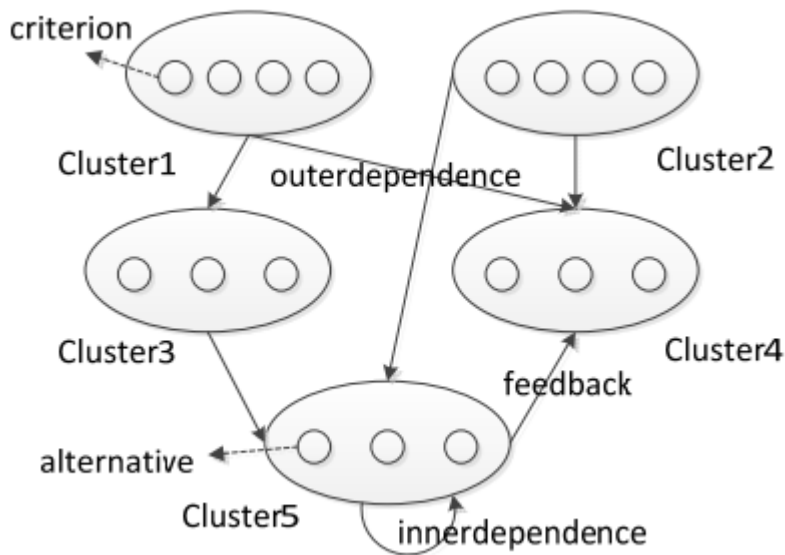


Figure 2.1: The Structure of ANP [16]

As in AHP method the nodes or clusters' pairwise comparison and calculation of local priorities are the same. Local priorities result from the pairwise comparison matrix' Eigen vector, found priorities are then structured in the super- matrix as column vectors. Un-weighted Super Matrix got after all the comparisons have done. From here and by squaring the matrix the alternatives ranking as the AHP method can got, the alternatives impact on the criteria importance should be take into consideration to transfer to the ANP, the matrix normalizes to get the weighted Super Matrix. Then

calculating the limit matrix to synthesize the model, which converges weighted super matrix, the final result is alternatives ranking. (More details in the algorithm implementation)

ANP algorithm involves the following steps [16]:

- Step 1 - Determination Problem: here the current problem is identified. Criteria of decision making problem sub-criteria and alternatives are determined.
- Step 2 - Determination of Relations with criteria: The interactions of the specified criteria with each other, the internal and external interactions of each criteria, and the existing feedback are associated with this step. The opinions of experts are taken and the literature about the current problem is searched.
- Step 3 - Performing Binary Comparisons between Criteria: As in the Analytic Hierarchy Process, pairwise comparison is made between each Criteria that is considered to be related to each other. These pairwise complements are aggregated into a resultant matrix.
- Step 4 - Checking Whether the Comparison Matrices Are Consistent: A consistency analysis is performed to see if the comparisons made in this step are meaningful. After the comparison values are given, the consistency rate symbolized as CR for each matrix is calculated.
- Step 5 - Generating Super Matrices in Order: In this step, inter-criterion evaluations are summarized under a large matrix under the name non-weighted super matrix. Then, multiplying the resultant super-matrix with the weighted values for corresponding clusters in the super-matrix. Taken to the $(2K+1)$ power, (K is arbitrary number)
- Step 6 - Determination and Selection of the Best Alternative: It is possible to make a comparison between the limit mathematical alternatives to see best alternative. Greatest value here represents the best alternative.

The AHP is a kind of network, it follows the up down model, where the work start from the goal cluster to the alternatives according to criteria, so its downward hierarchy, in contrast the ANP method is going in the two ways, where it not just study the criteria impact on alternatives, but

also it take on the consideration the alternative impact on the decision making .which represent the real case which faced in the real life as seen in the figure (2.2).

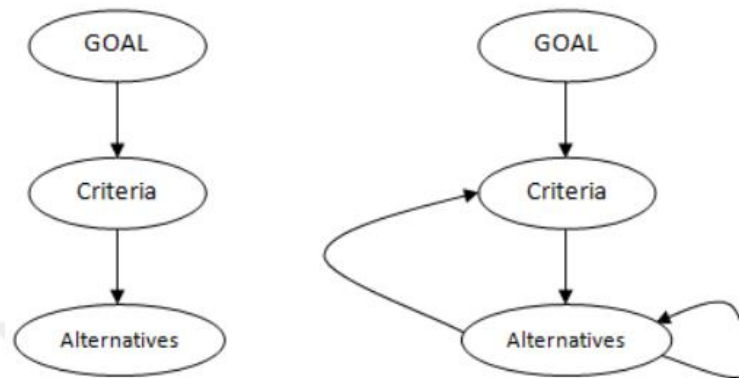


Figure 2.2: Difference in Structure between AHP (left) and ANP (Right). [17]

❖ Advantages

As mentioned above one of the ANP method's advantages is that the two directions links from the nodes to the clusters, which help to deal with complicated problem in the real life. Also it helps to understand our problem and the interactions between the elements better. [17]

❖ Disadvantages

1. In the ANP method a $n(n-1)/2$ pairwise comparisons are performed, which in turn make the comparison process more complicated and power consuming, for that a limited alternatives numbers and criteria should use; recommended number in cluster is less than five alternatives and criteria. [17]
2. Users tend to make the decision according the importance, it's hard to conceive the DM to make another pairwise comparison between items to reconsider their inputs, especially if the consistency index for the alternative ranking is too high. [17]
3. it's hard to apply the ANP method in Excel, so its needs a special software to implement the method. [17]

2.2.2 ELECTRE Method

ELECTRE means the elimination and selection that reflects the truth. Developed in 1966 by Roy and his friends. As a response to existing decision-making methods developed. In fact, it is not just a solution method is a debated philosophy. The main concept of the ELECTRE method; for each criterion is to use dual comparisons between alternatives. For each rating factor, it is based on binary superiority comparisons between alternative decision points. Where two alternatives are compared in a time and selects the one which is better in most criteria and not acceptably worse in other criteria. ELECTRE method a multi-purpose decision making technology used. [18]

The method steps are: [19]

- Step 1 - Preparation of Decision Matrix
- Step 2 - Calculate the normalized decision matrix.

It will be normalized using the following formula in eq (2.1).

$$x_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^n r_{ij}^2}} \quad i = 1, 2, \dots, M \quad j = 1, 2, \dots, n \quad (2.1)$$

The normalized matrix will be as following in eq (2.2)

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (2.2)$$

- Step 3 - Calculate weighted normalized decision matrix in eq (2.3).

$$Y_{ij} = \begin{bmatrix} w_1x_{11} & w_2x_{12} & \dots & w_nx_{1n} \\ w_1x_{21} & w_2x_{22} & \dots & w_nx_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ w_1x_{m1} & w_2x_{m2} & \dots & w_nx_{mn} \end{bmatrix} \quad (2.3)$$

- Step 4 -Determine the concordance and discordance set.

Y matrix used to determine the concordance sets. Decision points are compared with each other in terms of evaluation factors. Net weighted normalized matrix data is compared for every pair and results are evaluated as below: If alternative is better than or equal to other element of pair it is considered under concordance set and defined by C. sets is determined by the relationship shown in the form in equation (2.4).

$$C_{kl} = \{j, y_{kj} \geq y_{lj}\} \quad (2.4)$$

The formula is based on the size of the line elements relative to each other based on comparison.

If alternative is worse than the other element of the pair for relevant criteria it is considered under discordance set and defined by D. The discordance set can be calculate as following in equation (2.5).

$$D(p, q) = \{j, v_m < v_\sigma\} \quad (2.5)$$

- Step 5 -Calculate the concordance matrix.

Concordance matrix is the matrix generated by adding the values of weights of Concordance set elements in equation (2.6) and (2.7).

$$C_{pq} = \sum_{j^*} w_{j^*} \quad (2.6)$$

$$C = \begin{bmatrix} - & c_{12} & c_{13} & \dots & c_{1m} \\ c_{21} & - & c_{23} & \dots & c_{2m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ c_{m1} & c_{m2} & c_{m3} & \dots & - \end{bmatrix} \quad (2.7)$$

- Step 6 -Calculate the discordance matrix.

Discordance matrix is prepared by dividing discordance set members values to total value of whole set in equation (2.8) and (2.9).

$$D_{pq} = \frac{\left(\sum_{j^*} |v_{pj^*} - v_{qj^*}| \right)}{\left(\sum_j |v_{pj} - v_{qj}| \right)} \quad (2.8)$$

$$D = \begin{bmatrix} - & d_{12} & d_{13} & \dots & d_{1m} \\ d_{21} & - & d_{23} & \dots & d_{2m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ d_{m1} & d_{m2} & d_{m3} & \dots & - \end{bmatrix} \quad (2.9)$$

- Step 7- Make calculations of advantage Averages of concordance and discordance values are taken. In the Concordance matrix any C_{pq} value bigger than or equal to C average it is stated as yes. In the discordance matrix any value less than or equal to D average is stated as No.
- Step 8 -Calculate net concordance and discordance matrix

The best alternative is the one that dominates all the other alternatives in this manner.

To rank alternatives we calculate the net concordance and net discordance values, we use the following formulation in equation (2.10) and (2.11).

$$c_k = \sum_{\substack{l=1 \\ l \neq k}}^m c_{kl} - \sum_{\substack{l=1 \\ l \neq k}}^m c_{lk} \quad (2.10)$$

$$d_k = \sum_{\substack{l=1 \\ l \neq k}}^m d_{kl} - \sum_{\substack{l=1 \\ l \neq k}}^m d_{lk} \quad (2.11)$$

2.2.3 VIKOR Method

VIKOR (VIseKriterijumsa Optimizacija I Kompromisno Resenje) method has been proposed by Serafim Opricovic in 1998 to deal with very complex decision problems. The method used in many fields. [20]

Offers compelling solutions for problems with contradictory criteria, focusing on sorting and selecting alternatives. To reach final decisions. Best alternative solution is the most close solution to ideal, and best alternative is reaching agree on mutual acceptance. [21]

In the following the VIKOR method steps: [20]

- Step 1 - for each criterion ($i = 1, 2, \dots, n$), alternatives ($J = 1, 2, \dots, J$) we need to calculate the worst and the best alternative for each criterion:

If the i criterion represents utility we calculate as following in equation (2.12).

$$f_i^* = \max_i f_{ij} \quad f_i^- = \min_i f_{ij} \quad (2.12)$$

If the i criterion represents cost we calculate as following in equation (2.13).

$$f_i^* = \min_i f_{ij} \quad f_i^- = \max_i f_{ij} \quad (2.13)$$

- Step 2 - to each alternative, the following formula used to calculate the ideal value S_j (or benefit measure) and the negative value R_j (or regression measure) in equation (2.14).

$$S_j = \sum_{i=1}^n \omega_i \cdot \frac{(f_i^* - f_{ij})}{(f_i^* - f_i^-)} \quad R_j = \max_i \left[\omega_i \cdot \frac{(f_i^* - f_{ij})}{(f_i^* - f_i^-)} \right] \text{ where } j = 1, \dots, m \text{ and } i = 1, \dots, n \quad (2.14)$$

ω_i : Expresses criteria weights indicating relative importance. The sum of the weights will be equal to 1.

- Step 3 - Calculate the synergy value Q_j for each alternative using the following equations in eq (2.15).

$$Q_j = v \frac{(S_j - S^*)}{(S^- - S^*)} + (1-v) \frac{(R_j - R^*)}{(R^- - R^*)} \quad S^* = \min_j S_j \quad S^- = \max_j S_j \quad \text{where } j = 1, \dots, m \text{ and } i = 1, \dots, n \quad (2.15)$$

v expresses the weight of the maximum group benefit, $1-v$ the weight of personal regret [21]. v is generally taken as 0.5

- Step 4 - S , R and Q are sorted from small to large. S , R , and Q values are sorted in their own order to obtain three different orders

- Step 5 - The alternatives A (1) represents the best ordered solution in the order of decreasing order by the measured values S, R and Q and then Q (minimum).

When the proposed solution is proposed, the following conditions should fulfilled

- a. Acceptable advantage in equation (2.16).

$$Q(A^{(2)}) - Q(A^{(1)}) \leq DQ \text{ where } DQ = \frac{1}{m-1} \quad (2.16)$$

* A (2) indicates the second best alternative, m: the number of alternatives

- b. Acceptable stability when making a decision - the recommended alternative (1) should be ranked by S and / or R best.

If one of these two conditions cannot be met, then the agreed-upon common best solution set is proposed as follows:

Alternatives (1) and A (2) if condition (B) is not met. (A) are not fulfilled, the alternatives A (1), A (2)... A (m); A (m) is the maximum for the relationship to M, is determined in equation (2.17).

$$Q(A^{(M)}) - Q(A^{(1)}) < DQ \quad (2.17)$$

2.2.4 PROMETHEE Method

PROMETHEE (Preference Ranking Organization METHods Enrichment Evaluation) the method suggested by Jean-Pierre Brans in 1998.

considered from Partial Aggregation Methods, This method is able to evaluate a large set of alternatives based on a large set of criteria as a classification of these alternatives according to the priority and importance, and it was classified as one of the most efficient MCDM methods. The goal of the PROMETHEE method is to classify the alternatives from the most important to the least, so that each standard has a quantitative weight and each alternative has its own evaluation for this criterion; weights and ratings are used to calculate this compound preference index that determines how preferable one alternative is to another. [22]

The PROMETHEE built on three axioms: [22]

- 1) Examination: if two alternatives have the same estimation for each criterion, then the decision maker sees the neutrality between these alternatives.
- 2) Cohesion: if alternative a is better than alternative b for each criterion, then a is better than b in the final result
- 3) Non-Redondance: a criterion is non-redundant if deleting it prevented the criteria group from achieving the previous axioms

The procedure of the PROMETHEE method consists of several steps: [22]

- Step 1 - The pairwise comparison for each two alternatives according to each criterion:

In general there are four relations types between alternatives: [22]

- ✓ Indifference: there are clear reasons explain the neutrality between two alternatives.
- ✓ Preference Stricte: there are clear reasons explain the superiority for one alternative comparing the other.
- ✓ Poor preference: there are clear reasons eliminate the superiority for one alternative comparing the other.
- ✓ Incomparability: there are none of the previous relation exists we take this relation.

The evaluation table is represent the main base in PROMETHEE method, where it contains the alternatives, criteria, weights, thresholds, as in the following table (2.1) (the table is taken from the main interface for “visual PROMETHEE” application) : [23]

Table 2.1: The Evaluation Table [23]

Criteria	G1	G2	G n
preferences				
Weights	W1	W2	W n
Preference function				
Thresholds	P1	P2	Pn
	Q1	Q2	Qn
	S1	S2	Sn
Alternatives				
A1	G1 (a1)	G2 (a1)		Gn (a1)
Am	G1 (am)	G2 (am)		Gn (am)

Criteria and alternatives discussed previously

Weights: are the importance of each criterion according others

Thresholds: determined by the decision makers, where there are three types: P, Q, and S

- ✓ Indifference threshold “Q”: it’s the max value that keep the decision maker neutral from choosing one between two alternatives.
- ✓ Preference threshold “P”: the min value that make the decision maker prefer one alternative between two.

$$\min |d_j (a, b)| \leq q \leq p \leq \max |d_j (a, b)|$$

Where:

$$\text{si } d_j(a, b) < Q \Rightarrow P_j(a, b) = 0$$

$$\text{si } d_j(a, b) > P \Rightarrow P_j(a, b) = 1$$

$d_j(a, b)$ represents the difference between two values a, b according to criterion g where:

$$d_j(a, b) = g_j(a) - g_j(b)$$

$g_j(a)$: represent the estimation for the alternative a according to criterion g

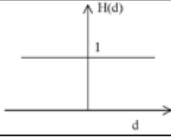
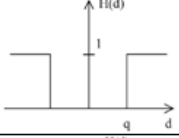
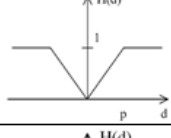
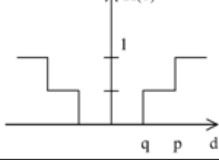
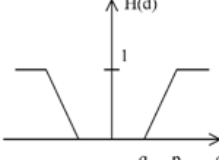
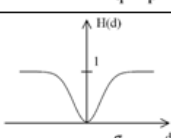
$g_j(b)$: represent the estimation for the alternative b according to criterion g

$d_j(a, b)$: the difference function which represent the preference between a and b according to g

- ✓ Gaussian threshold “S”: If the difference between evaluating two alternatives is greater than this threshold, the decision-maker avoids the alternative that contributed to this neutrality

To select the preference function, there are six criteria as seen in the following table (22) shows [22].

Table 2.2: Preference Functions Criteria [22]

Type of generalized criteria	Analytical definition	Shape	Parameters to define
Type I. Usual criterion	$H(d) = \begin{cases} 0, & d = 0; \\ 1, & d > 0. \end{cases}$		--
Type II. Quasi-criterion	$H(d) = \begin{cases} 0, & d \leq q; \\ 1, & \text{otherwise.} \end{cases}$		q
Type III. Criterion with linear preference	$H(d) = \begin{cases} \frac{ d }{p}, & d \leq p; \\ 1, & d > 0. \end{cases}$		p
Type IV. Level-criterion	$H(d) = \begin{cases} 1, & d \leq q; \\ 1/2, & q < d \leq p; \\ 1, & \text{otherwise.} \end{cases}$		q, p
Type V. Criterion with linear preference and indifference area	$H(d) = \begin{cases} 1, & d \leq q; \\ \frac{ d - q}{p - q}, & q < d \leq p; \\ 1, & \text{otherwise.} \end{cases}$		q, p
Type VI. Gaussian criterion	$H(d) = 1 - \exp\left(-\frac{d^2}{2\sigma^2}\right)$		σ

- Step 2 - For each couple of actions a, b E- K, a preference index π should define for a with regard to b over all the criteria. Suppose every criterion has been identified as being of one of the six types considered so that the preference functions $P_h(a,b)$ have been defined for each $h = 1, 2, ..k$.

Supposing that all the criteria have the same importance. If it is not the case, one can introduce a weighted preference index. As following in equation (2.18). [24]

$$\pi(a, b) = \frac{1}{k} \sum_{h=1}^k P_h(a, b) \quad (2.18)$$

- Step 3- calculate the flows: [24]

- ✓ Outgoing flow in equation (2.19).

$$phi^+(a) = \frac{1}{n-1} \sum_{x \in A} \pi(a, x) \quad (2.19)$$

- ✓ Incoming flow in equation (2.20).

$$phi^-(a) = \frac{1}{n-1} \sum_{x \in A} \pi(x, a) \quad (2.20)$$

- ✓ Net flow in equation (2.21).

$$phi(a) = phi^+(a) - phi^-(a) \quad (2.21)$$

- Step 4- ranking alternatives [22]

- ✓ PROMETHEE I ranking: alternative a preferred if $phi^+(a)$ is large and $phi^-(a)$ is small
- ✓ PROMETHEE II ranking: alternative a preferred on b if $phi(a) > phi(b)$

3. METHODOLOGY

3.1 RESEARCH GOAL

In this research problem, there are eight features groups, which regard as alternatives should be ranked to choose the most important between them to use in the our classification problem to improve the accuracy prediction, these features will ranked according to six famous measures which will regard as criteria.

Kaur et al. research, who used the "AHP-TOPSIS" which is one of the MCDM method to rank the feature groups [7]; this work will continue Kaur et al. research by using other MCDM methods (ANP, ELECTRE, PROMETHEE and VIKOR), each one of these method will be implemented on the Kaur et al. research dataset mentioned in table (1.1), and the result of them will be compared to show the most features group agreed by them

3.1.1 Alternatives

Seven features groups will be used as alternatives, as following:

Char n-gram "c", Unigram "U", Bigram "B", SPATIUM-L1 "L1", TFIDF "T", Stylometric "S", NMF "N".

3.1.2 Criteria

Six measures will be used as criteria, as following:

F-score "F", Recall "R", Precision "P", False Acceptance Rate "FAR", False Rejection Rate "FRR", Accuracy "A"

We have the following table (3.1) from the Kaur study:

Table 3.1: Accuracy Table [7]

	TRUE RATE ACCURACY			FALSE RATE ACCURACY		Co-efficient of Variance
	<i>F</i>	<i>R</i>	<i>P</i>	<i>FAR</i>	<i>FRR</i>	CV
C	87.50	84.00	91.30	7.41	16.00	88.46
U	90.57	96.00	85.71	14.81	4.00	90.38
B	86.27	84.61	88.00	10.71	15.38	87.04
L1	79.36	100.00	65.78	48.14	0.00	75.00
T	83.02	88.00	78.57	22.22	12.00	82.69
S	77.55	76.02	79.16	18.52	24.02	78.85
N	84.74	100.0	73.53	33.33	0.00	82.69

3.2 PROBLEM SOLVING

3.2.1 Solution of Problem with ANP Method

Super Decisions V2 software was used to construct the relationships of internal networks among the criteria and alternatives, internal dependencies and external dependencies, to make binary comparisons and to calculate weights.

The following Figure shows the network structure defined between the criteria and alternatives set out for the problem of selecting feature group.

An internal dependency between F-score “F”, Recall “R”, Precision “P”, within the TRUE RATE ACCURACY set is defined.

An internal dependency is defined between the False Acceptance Rate “FAR”, False Rejection Rate “FRR”, within the False Rate Accuracy.

There is a relationship between each criterion and all other criteria in other clusters.

The figure (3.1) shows the network relations between the alternatives and the criteria:

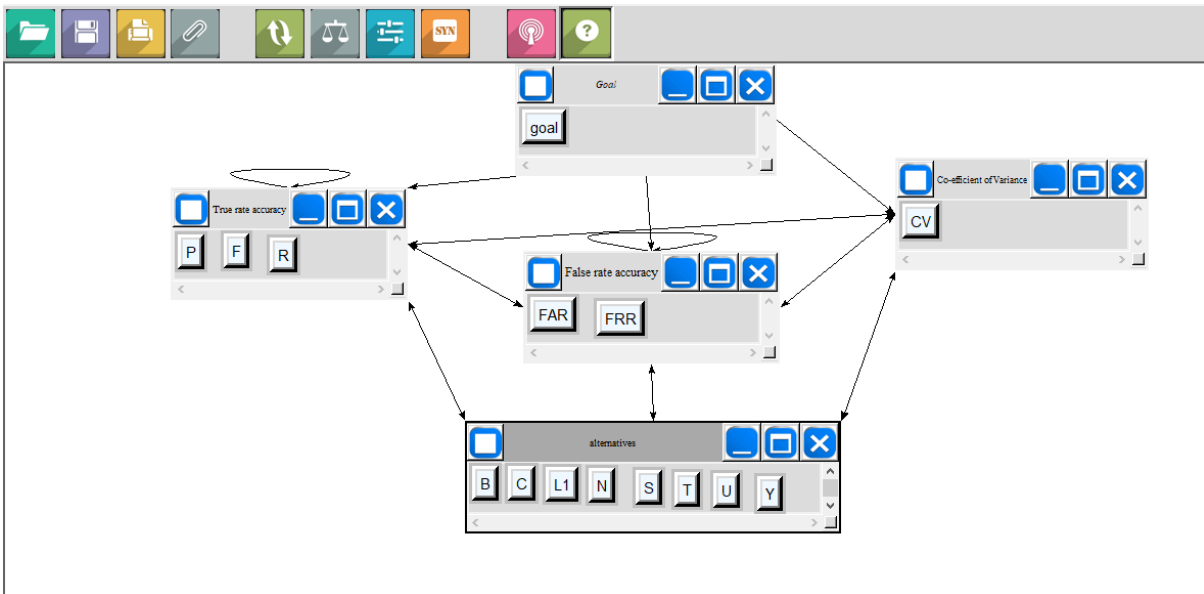


Figure 3.1: Network Structure Defined Among the Criteria and Alternatives

The figure (3.2) shows the output of the super reconciliation program screen, where the solution priorities of the alternatives identified for the feature group selection problem are based on the ANP method.

New synthesis for: Super Decisions Ma... _ □ ×

Here are the overall synthesized priorities for the alternatives. You synthesized from the network Super Decisions Main Window: 1.sdmod

Name	Graphic	Ideals	Normals	Raw
B	<div style="width: 82.8%; background-color: blue;"></div>	0.828388	0.138017	0.029575
C	<div style="width: 80.5%; background-color: blue;"></div>	0.805201	0.134154	0.028747
L1	<div style="width: 91.6%; background-color: blue;"></div>	0.916128	0.152635	0.032708
N	<div style="width: 80.2%; background-color: blue;"></div>	0.802255	0.133663	0.028642
S	<div style="width: 100%; background-color: blue;"></div>	1.000000	0.166609	0.035702
T	<div style="width: 89.0%; background-color: blue;"></div>	0.890008	0.148283	0.031775
U	<div style="width: 76.0%; background-color: blue;"></div>	0.760100	0.126639	0.027137

Figure 3.2: Super Decision Screen Display of Priorities of Alternatives

The following table (3.2) shows the result of the ANP method application, where features group are ranked according to their ability to achieve the goal, Stylometric feature group “S”takes the first order.

Table 3.2: ANP Ranking Result

Graphic	Alternatives	Total	Normal	Ideal	Ranking
	B	0.0296	0.1380	0.8284	4
	C	0.0287	0.1342	0.8052	5
	L1	0.0327	0.1526	0.9161	2
	N	0.0286	0.1337	0.8023	6
	S	0.0357	0.1666	1.0000	1
	T	0.0318	0.1483	0.8900	3
	U	0.0271	0.1266	0.7601	7

3.2.2 Solving the problem with ELECTRE method

After normalizing the base matrix by applying the following formula in eq (3.1).

$$x_{ij} = \frac{a_{ij}}{\sqrt{\sum_{k=1}^m a_{kj}^2}} \quad (3.1)$$

We get the R matrix table (3.3).

Table 3.3: Normalized Original Matrix

	<i>F</i>	<i>R</i>	<i>P</i>	<i>FAR</i>	<i>FRR</i>	<i>CV</i>
C	0.3618	0.3901	0.3284	0.1085	0.4448	0.3674
U	0.3745	0.3662	0.3753	0.2168	0.1112	0.3754
B	0.3567	0.3760	0.3308	0.1568	0.4275	0.3615
L1	0.3281	0.2810	0.3910	0.7046	0.0000	0.3115
T	0.3432	0.3357	0.3440	0.3252	0.3336	0.3435
S	0.3206	0.3357	0.2972	0.2711	0.6677	0.3275
N	0.3504	0.3142	0.3910	0.4878	0.0000	0.3435

The weights of each criteria are then multiplied by the normalized matrix (R matrix) as follows in equation (3.2).

$$V_{ij}=RxW = \begin{bmatrix} r_{11} \cdot W_1 & r_{12} \cdot W_2 \dots & r_{1n} \cdot W_n \\ r_{21} \cdot W_1 & r_{22} \cdot W_2 \dots & r_{2n} \cdot W_n \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ r_{m1} \cdot W_1 & r_{m2} \cdot W_2 \dots & r_{mn} \cdot W_n \end{bmatrix} \quad (3.2)$$

The V matrix has been got table (3.4).

Table 3.4: Weighted Normalized Matrix

	<i>F</i>	<i>R</i>	<i>P</i>	<i>FAR</i>	<i>FRR</i>	<i>CV</i>
C	0.1197	0.0351	0.0761	0.0211	0.0476	0.0231
U	0.1239	0.0402	0.0714	0.0423	0.0119	0.0236
B	0.1181	0.0354	0.0733	0.0306	0.0457	0.0228
L1	0.1086	0.0418	0.0548	0.1374	0.0000	0.0196
T	0.1136	0.0368	0.0655	0.0634	0.0357	0.0216
S	0.1061	0.0318	0.0660	0.0529	0.0714	0.0206
N	0.1160	0.0418	0.0613	0.0951	0.0000	0.0216

By applying the next formula on the V matrix in equation (3.3).

$$C_{ab} = \{j | x_{aj} \geq x_{bj}\} \quad (3.3)$$

We get:

$$C_{12} = \{3,5\}$$

$$C_{34} = \{1,2,4\}$$

$$C_{21} = \{1,2,4,6\}$$

$$C_{43} = \{3,5,6\}$$

$$C_{13} = \{1,3,5,6\}$$

$$C_{35} = \{1,3,5,6\}$$

$$C_{31} = \{2,4\}$$

$$C_{53} = \{2,4\}$$

$$C_{14} = \{1,3,5,6\}$$

$$C_{36} = \{1,2,3,6\}$$

$$C_{41} = \{2,4\}$$

$$C_{63} = \{4,5\}$$

C15 = {1,3,5,6}	C37 = {1,3,5,6}
C51 = {2,4}	C73 = {2,4}
C16 = {1,2,3,6}	
C61 = {4,5}	C45 = {2,4}
C17 = {1,3,5,6}	C54 = {1,3,5,6}
C71 = {2,4}	C46 = {1,2,4}
	C64 = {3,5,6}
C23 = {1,2,4,6}	C47 = {2,4,5}
C32 = {3,5}	C74 = {1,3,6}
C24 = {1,3,5,6}	
C42 = {2,4}	C56 = {1,2,4,6}
C25 = {1,2,3,6}	C65 = {3,5}
C52 = {4,5}	C57 = {3,5}
C26 = {2,3,6}	C75 = {1,2,4,6}
C62 = {1,4,5}	
C27 = {1,3,5,6}	C67 = {3,5}
C72 = {2,4}	C76 = {1,2,4,6}

And by applying the next formula on the V matrix in eq (3.4).

$$D_{ab} = \{j | x_{aj} < x_{bj}\} = j - C_{ab} \quad (3.4)$$

We get:

$$D21 = \{3,5\}$$

$$D43 = \{1,2,4\}$$

$$D12 = \{1,2,4,6\}$$

$$D34 = \{3,5,6\}$$

$$D31 = \{1,3,5,6\}$$

$$D53 = \{1,3,5,6\}$$

$$D13 = \{2,4\}$$

$$D35 = \{2,4\}$$

$$D41 = \{1,3,5,6\}$$

$$D63 = \{1,2,3,6\}$$

$$D14 = \{2,4\}$$

$$D36 = \{4,5\}$$

$$D51 = \{1,3,5,6\}$$

$$D73 = \{1,3,5,6\}$$

$$D15 = \{2,4\}$$

$$D37 = \{2,4\}$$

$$D61 = \{1,2,3,6\}$$

$$D16 = \{4,5\}$$

$$D54 = \{2,4\}$$

$$D71 = \{1,3,5,6\}$$

$$D45 = \{1,3,5,6\}$$

$$D17 = \{2,4\}$$

$$D64 = \{1,2,4\}$$

$$D46 = \{3,5,6\}$$

$$D32 = \{1,2,4,6\}$$

$$D74 = \{2,4,5\}$$

$$D23 = \{3,5\}$$

$$D47 = \{1,3,6\}$$

$$D42 = \{1,3,5,6\}$$

$$D24 = \{2,4\}$$

$$D65 = \{1,2,4,6\}$$

$$D52 = \{1,2,3,6\}$$

$$D56 = \{3,5\}$$

$$D25 = \{4,5\}$$

$$D75 = \{3,5\}$$

$$D62 = \{2,3,6\}$$

$$D57 = \{1,2,4,6\}$$

$$D26 = \{1,4,5\}$$

$$D72 = \{1,3,5,6\}$$

$$D76 = \{3,5\}$$

$$D27 = \{2,4\}$$

$$D67 = \{1,2,4,6\}$$

Next, we use the previous result to construct the matrix C; the elements are calculated as follows in equation (3.5) table (3.5).

$$c_{ab} = \sum_{j \in C_{ab}} w_j \quad (3.5)$$

weights	0.331	0.195	0.107	0.195	0.107	0.063
---------	-------	-------	-------	-------	-------	-------

Table 3.5: C matrix

0	0.577634	1	1	1	0.46526	0.120181
1	0	1	1	1	1	0.11529
0.200575	0.337477	0	1	1	1	1
0.419376	0.174726	0.438667	0	0.494247	0.865737	0.218167
0.288339	0.459246	0.313501	1	0	0.288336	0.867231
0.46526	0.211954	0.491124	1	0.288336	0	1
0.658996	0.230578	0	1	1	1	0

And then we form the matrix D, which calculates these elements as follows in equation (3.6) table (3.6).

$$d_{ab} = \frac{\max_{j \in D_{ab}} |v_{aj} - v_{bj}|}{\max_{j \in J, m, n \in I} |v_{mj} - v_{nj}|} \quad (3.6)$$

Table 3.6: D matrix

0	0.214	0.608	0.608	0.608	0.696	0.608
0.784	0	0.784	0.608	0.696	0.696	0.608
0.39	0.214	0	0.608	0.608	0.696	0.608
0.39	0.39	0.39	0	0.39	0.721	0.497
0.39	0.302	0.39	0.608	0	0.784	0.277
0.302	0.302	0.302	0.277	0.214	0	0.214
0.39	0.39	0.39	0.803	0.784	0.784	0

And next, we need to calculate \bar{c} from the following formula in equation (3.7).

$$\bar{c} = \sum_{a=1}^m \sum_{b=1}^m c(a, b) / m(m - 1) \quad (3.7)$$

Which equal to (0.416416), so we compare each element of C matrix with this value to get E matrix which its elements calculated by the following formulation in equation (3.8) (3.9).

$$e(a, b) = 0; \quad \text{if } (a, b) < \bar{c} \quad (3.8)$$

$$e(a_a b) = 1; \quad \text{if } (a_a b) \geq \bar{c} \quad (3.9)$$

So, we get E matrix as seen in the following table (3.7):

Table 3.7: E matrix

0	1	1	1	1	0	0
1	0	1	1	1	1	0
0	0	0	1	1	1	1
0	0	0	0	1	1	0
0	0	0	1	0	0	1
0	0	1	1	0	0	1
1	0	0	1	1	1	0

And then -in the same way-, we form the matrix F from matrix D as seen in the following table (3.8):

Table 3.8: F matrix

0	0	1	1	1	1	1
1	0	1	1	1	1	1
1	0	0	1	1	1	1
1	1	1	0	1	1	1
1	0	1	1	0	1	0
0	0	0	0	0	0	0
1	1	1	1	1	1	0

And in the last step we use the C and D matrices to derive the final result according to these two formulas in equation (3.10) and (3.11). table (3.9):

$$c_k = \sum_{\substack{l=1 \\ l \neq k}}^m c_{kl} - \sum_{\substack{l=1 \\ l \neq k}}^m c_{lk} \quad (3.10)$$

$$d_k = \sum_{\substack{l=1 \\ l \neq k}}^m d_{kl} - \sum_{\substack{l=1 \\ l \neq k}}^m d_{lk} \quad (3.11)$$

Table 3.9: Ck and Dk values

C1	1.13053	D1	0.696
C2	3.123677	D2	2.364
C3	1.294759	D3	0.26
C4	-3.38908	D4	-0.734
C5	-1.56593	D5	-0.549
C6	-1.16266	D6	-2.766
C7	0.568704	D7	0.729

The final result is shown in table (3.10):

Table 3.10: ELECTRE Ranking Result

	Top Value	Top Ranking	Lowest Value	Bottom Value Order
C	1.13053	4	0.696	4
U	3.123677	1	2.364	1
B	1.294759	7	0.26	5
L1	-3.38908	3	-0.734	3
T	-1.56593	2	-0.549	2
S	-1.16266	5	-2.766	6
N	0.568704	6	0.729	7

3.2.3 Solution of Problem with VIKOR Method

For each criterion f^+_i and f^-_i values calculated as seen in the following table (3.11):

Table 3.11: F+I and F- I' values

criteria	f^+_i	f^-_i
F	90.57	77.55
R	100	76.02
P	91.3	65.78
FAR	48.14	7.41
FRR	24.02	0
CV	90.38	75

For each alternative S_j and R_j values calculated in equation (3.12) and (3.13) table (3.12):.

$$S_j = \sum_{k=1}^n w_i (f_i^+ - f_{ij}) / (f_i^+ - f_i^-) \quad (3.12)$$

$$R_j = \max [w_i (f_i^+ - f_{jj}) / (f_i^+ - f_i^-)] \quad (3.13)$$

Table 3.12: S_j and R_j Values

Alternatives	S _j	R _j
C	0.446746	0.195
U	0.304718	0.159572
B	0.479671	0.179201
L1	0.561985	0.284985
T	0.552035	0.191939
S	0.765939	0.331
N	0.432123	0.148213

Then Q_j values by using the following formula calculated in equation (3.14) table (3.13):.

$$Q_j = \frac{v(S_j - S^+)}{(S^- - S^+)} + \frac{(1 - v)(R_j - R^+)}{(R^- - R^+)} \quad (3.14)$$

Table 3.13: Q_j ' Values

Alternatives	Q
C	0.281952
U	0.031071
B	0.274428
L1	0.653029
T	0.387721
S	1
N	0.138117

After that we rank the alternatives as seen in the following table (3.14):

Table 3.14: VIKOR ranking result

Alternatives	Q	ranking
C	0.281952	3
U	0.031071	1
B	0.274428	4
L1	0.653029	6
T	0.387721	5
S	1	7
N	0.138117	2

3.2.4 Solution of Problem with PROMETHEE Method

The "Visual PROMETHEE Academic" program used to build the solution, the program interface as in the figure (3.3). Eight columns added representing each criterion, in the cluster line each two criteria together have been combined which represented by the same shape; in the Min/Max line Min selected if want to minimize this criterion or Max if want to maximize it; the weight extracted from the weight matrix which used previously; the appropriate preference function selected for each criterion in the preference function line; threshold selected according to the preference function selected ; the other statics measure extracted from the original matrix.

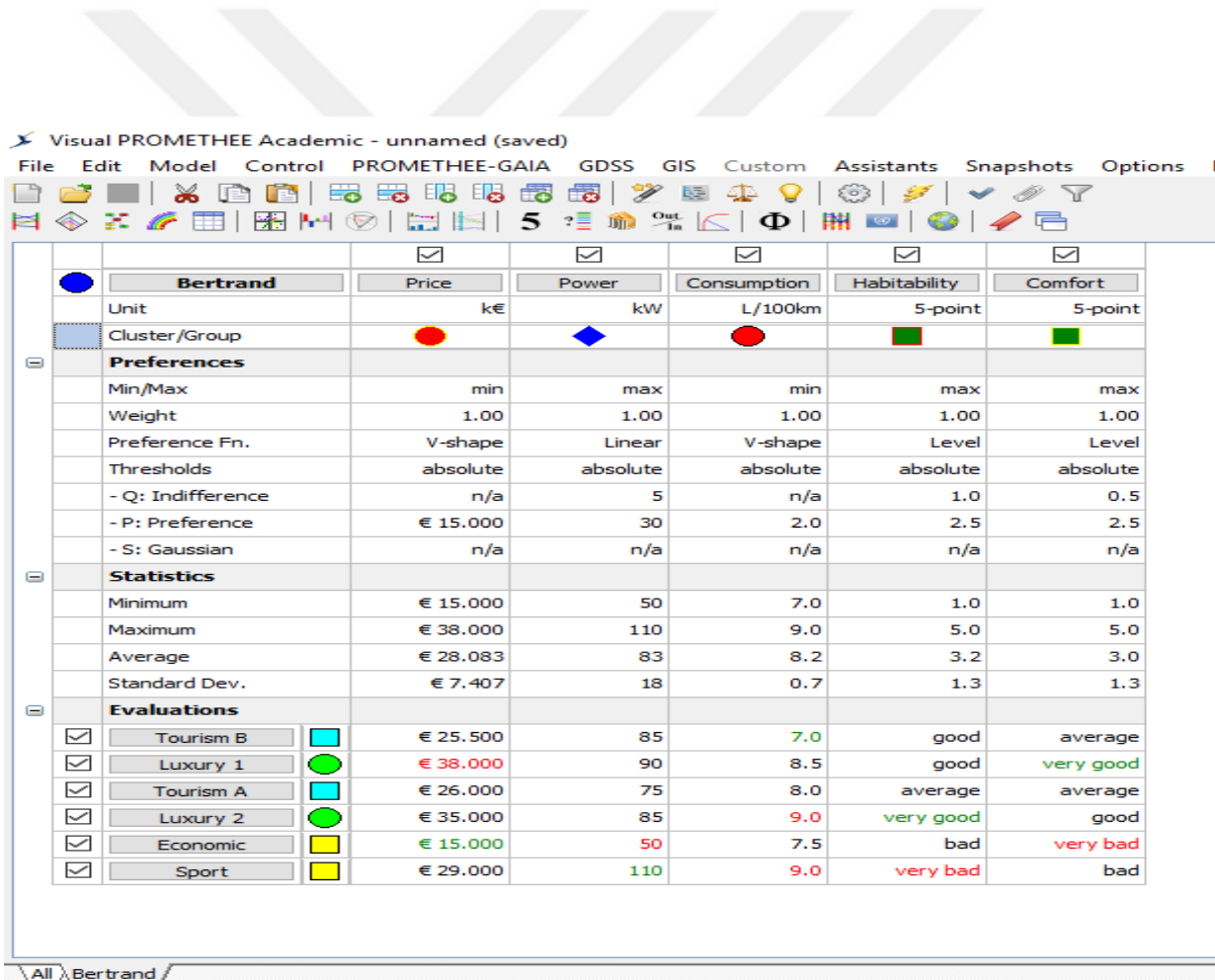


Figure 3.3: Visual PROMETHEE Academic" Program Interface

In the following table (3.15) the alternatives ranking result:

Table 3.15: PROMETHEE Ranking Result

Alternatives	Total	Normal	Ideal	Ranking
U	0.3402	0.6679	0.3295	1
N	0.3251	0.6112	0.2862	2
L1	0.1747	0.5567	0.3820	3
T	-0.0011	0.4702	0.4713	4
B	-0.0539	0.4289	0.4828	5
C	-0.1216	0.4001	0.5217	6
S	-0.6634	0.1437	0.8071	7

3.3 COMPARISON

Due to the previous implementation for the four MCDM methods to choose the most appropriate feature group to our case study, a different estimation has been got for the different feature groups, for that there is a need to compare the methods implementation' results to choose the ranking which most methods agreed on it. The following table contains the feature group ranking according to all the used MCDM algorithms, the table shows a contradictory between the ranks, we need such a process to extract the features ranking from the table (3.16).

Table 3.15: Methods Ranking Results

Features group names	ANP	ELECTRE	VIKOR	PROMETHEE
C	5	4	3	6
U	7	1	1	1
B	4	7	4	5
L1	2	3	6	3
T	3	2	5	4
S	1	5	7	7
N	6	6	2	2

The previous implementation shows the ranking for the features group according to the MCDM algorithms, where it's seen that more one algorithms are agreed on the priority of the Unigram "U" features took the first priority according to three algorithms (PROMETHEE, VICOR, and ELECTRE).

To extract the ranking for all the used MCDM algorithms, at first the feature which has the greater compatibility among the algorithms will be selected here is "U" feature, the second feature is "N" feature, the third feature is "L1", the fourth feature is "B", the fifth features is "T", the sixth is "C" and the seventh is "S", the following table shows the feature ranking for all the features which have the greater compatibility among the algorithms as seen in the following table (3.17).

Comparing with Kaur et al. research, it's clear that their ranking result is more near to the result of VIKOR method, where the ranking of the features are near, where the preferences order for Kaur et al.' works was as following: Folksonomy features – Unigram - Char n-gram - tfidf - Stylometric - NMF - SPATIUM-L1 . Where they gave more importance to the Folksonomy features, in contract to our research which gave more important to the Unigram feature as three algorithms: ELECTRE, VIKOR, PROMETHEE agreed on the first preference for the Unigram feature.

Table 3.17: Feature Ranking after Filtering

Features group names	ANP	ELECTRE	VIKOR	PROMETHEE	Kaur et al. [7]
C	5	4	3	6	3
U	7	1	1	1	2
B	4	7	4	5	1
L1	2	3	6	3	7
T	3	2	5	4	4
S	1	5	7	7	5
N	6	6	2	2	6

4. CONCLUSION

This research discussed the ability to select one from the exist textual feature groups to use them in extracting the author style from the tweets, in order to detect the tweet authorship; there are a lot of textual features could be extracted, in this study a comparison done between seven textual feature alternatives; differentiated according to six criteria represented the measures which detect the prediction accuracy for each alternative.

Within the scope of the study, a textual features selection problem is one of the MCDM problem; four MCDM methods used to discover the most appropriate feature (ANP, ELECTRE, PROMETHEE and VICOR methods), all the methods have been explained from the theoretical side and implemented on our case using a different tools.

Consequently, to compare between the ranking results, all the ranking results for the four algorithms are aggregated in one table to extract the feature which has the greater compatibility among the algorithms, the Unigram “U” features took the first priority according to three algorithms (PROMETHEE, VICOR, and ELECTRE), which make it the preferred one among the other features. And the other features are ranked according to their appropriate to our problem; which take off the ambiguity facing the researcher in sufficient features selecting.

REFERENCES

- [1] <https://www.nltk.org/book/ch06.html>
- [2] Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. Page 110 . ISBN 978-0-262-01243-0.
- [3] Xuan ,P.; Guo ,M. Z.; Wang ,J.; Liu ,X. Y.; Liu ,Y. (2011). "Genetic algorithm-based efficient feature selection for classification of pre-miRNAs". Genetics and Molecular Research. 10 (2): 588–603. PMID 21491369. doi:10.4238/vol10-2gmr969.
- [4] Stamatatos, E. (2009), "A survey of Modern authorship attribution methods", Journal of the American Society for Information Science and Technology, 538-556.
- [5] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, et al. (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man. Scientific reports, 5:10312.
- [6] Guyon, Isabelle; Elisseeff, André (2003). "An Introduction to Variable and Feature Selection". JMLR. 3.
- [7] Kaur, R., Singh, S., & Kumar, H. (2018). Auth Com: Authorship verification and compromised account detection in online social networks using AHP-TOPSIS embedded profiling based technique. Expert Systems with Applications, 113, 397-414.
- [8] Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010, July). Detecting spammers on twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS) (Vol. 6, No. 2010, p. 12).
- [9] Criado, N., Rashid, A., & Leite, L. (2016). Flash mobs, Arab Spring and protest movements: Can we analyse group identities in online conversations?. Expert Systems with Applications, 62, 212-224.

- [10] Ma, X., Feng, Y., Qu, Y., & Yu, Y. (2018). Attribute Selection Method based on Objective Data and Subjective Preferences in MCDM. *International Journal of Computers, Communications & Control*, 13(3).
- [11] Peng, Y., Kou, G., Ergu, D., Wu, W., & Shi, Y. (2012). An integrated feature selection and classification scheme. *Studies in Informatics and Control*, ISSN, 1220-1766.
- [12] Singh, R., Kumar, H., & Singla, R. K. (2014). TOPSIS based multi-criteria decision making of feature selection techniques for network traffic dataset. *International Journal of Engineering and Technology*, 5(6), 4598-4604.
- [13] Jahan, A., & Edwards, K. L. (2013). *Multi-criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design*. Butterworth-Heinemann.
- [14] Fülöp, János. (2005). "Introduction to decision making methods." BDEI-3 workshop, Washington.
- [15] husam A.M. abowatfa. 2015. Using Analytic Hierarchy Process for Prioritizing Industrial Sector in Palestine to Achieve Sustainable Development
- [16] Saaty, Thomas L. "Making and validating complex decisions with the AHP/ANP." *Journal of Systems Science and Systems Engineering* 14.1 (2005): 1-36.
- [17] Brestovac, Goran, and Robi Grgurina. "Applying multi-criteria decision analysis methods in embedded systems design." (2013).
- [18] DOĞU, EREN, and ZEYNEP FİLİZ. Bayesian aggregation methods for analytic hierarchy process and analytic network process in group decision making. Diss. DEÜ Fen Bilimleri Enstitüsü, 2012.
- [19] Yücel, M. Gökhan, and Ali Görener. "Decision Making for Company Acquisition by ELECTRE Method." *International Journal of Supply Chain Management* 5.1 (2016): 75-83.
- [20] Brestovac, Goran, and Robi Grgurina. "Applying multi-criteria decision analysis methods in embedded systems design." (2013).

- [21] Opricovic, Serafim, and Gwo-Hshiong Tzeng. "Extended VIKOR method in comparison with outranking methods." *European journal of operational research* 178.2 (2007): 514-529.
- [22] CHENIA, S. (2015) applying PROMETHEE in selecting employee (CRSTRA) Doctoral dissertation, gestion
- [23] PROMETHEE official website,2011 www.promethee-gaia.net/software.html
- [24] J. P. Brans and Ph. Vincke . A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making) .Source: *Management Science*, Vol. 31, No. 6 (Jun., 1985), pp. 647-656