International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Topic-based Opinion Leaders in Social Networks by Content and User Information

Berna Altınel Girgin[1*], Mustafa Abdullah Hakkoz[2], Emre Barkın Bozdağ[3], Murat Can Ganiz[4]

Submitted: 17/08/2020 Accepted : 17/12/2020

Abstract: Social media is like a revolution since it has changed many things in people's lifestyles by bringing new trends in communication, shopping, working…etc. Inspired by the importance of social media, we propose Opinion Leader Detection (OLED) system in this paper. OLED has two main parts. In the first part, the tweets were labeled with their categories by semantic kernels for topic-based analysis. We run these semantic kernels and their variants with SVM in our experiment environment. After LDA and these semantic classifiers, obtain category information for each tweet in the dataset. OLED's second part attempts to detect whether the users are opinion leaders in their category. Then, the leadership scores are calculated with the formula generated and opinion leaders are determined in each category. The purpose of OLED is to find the opinion leaders for each category. In other words, OLED aims to detect opinion leaders for different topics such as *Economy, Culture-Art, Politics, Sports and Technology.* We performed our experiments on a real data collection gathered from Twitter that includes 17,234,924 tweets and 38,727 users. The language of the dataset is Turkish. Users with highest scores are stated as opinion leaders. In order to evaluate OLED's performance, we also run PageRank algorithm on the same dataset. We also compare our study to one of the existing studies in the literature. The experimental results show that our framework OLED generates remarkable performance in compare to PageRank algorithm nearly in all topics and all selected top number of opinion leaders.

Keywords: *social network analysis, opinion leader detection, flow of influence, PageRank algorithm, semantic kernels*.

1. Introduction

As the usage of social networks has increased, the way people communicate has also dramatically changed. Interactions between different people are mainly driven by network connection mechanisms on social networks such as following or friending. These create huge social networks consisting hundreds of millions of users.

Within this network structure, certain groups of users frequently produce usually popular textual or visual content to share their experiences and thoughts with other users; others may have a more passive stance; usually consume the available content. These content producers usually have more connections than others do. These people, also known as opinion leaders, can access thousands of people with the contents they shared, and more importantly, they can influence an emotions and thoughts of a significantly large group of people due to the flow in these large social networks. This phenomenon has attracted the attention of both researchers and marketing industry. This has of course many practical applications in the marketing domain. Opinion leaders and influencers are usually employed by companies to promote their products and services. In fact, majority of the marketing budget is moving from traditional media to social media. [29]. Therefore, it is of great importance to analyze and identify these opinion leaders and influencers in the social networks especially in the context of different topics.

The *Opinion Leaders Detection (OLED)* system that is proposed in this paper consists of two main parts. A social network is created by collecting user information including tweets shared within a specific period. In the first part, the tweets were labeled with their categories using semantic kernels for topic-based analysis [14, 15]. After the category information is obtained, the second part attempts to detect whether a user is an opinion leader in their category or not. For each user, a leadership score is calculated with a formula we propose and opinion leaders are determined for each category. So our contribution is two folded, first we propose the use of a novel method for topic detection in this domain. This is important because especially for marketing purposes influencers should be detected on a certain topic. Secondly, we form a novel opinion leader calculation system.

The remainder of the paper is arranged as follows: Section 2 includes related work and background information. The proposed text classification algorithms and opinion leader detection method is explained in Section 3. The experimental setup and the corresponding experiment results including some discussion points are presented in Section 4. Finally, Section 5 gives concluding remarks and future directions.

2. Related Work

Various methods have been suggested in the literature for opinion leader detection and flow of influence. These methods can be grouped into five categories, namely; 1.) Diffusion based approaches, 2.) Graph-based approaches, 3.) Statistical and stochastic approaches, 4.) Page-Rank based approaches, 5.)

[1] *Computer Eng. Department, Marmara University, Istanbul, Turkey*
ORCID ID : 0000-0001-5544-0925
[2] *Computer Eng. Department, Marmara University, Istanbul, Turkey*
ORCID ID : 0000-0002-2963-8513
[3] *Computer Eng. Department, Marmara University, Istanbul, Turkey*
ORCID ID: 0000-0001-6961-9248
[4] *Computer Eng. Department, Marmara University, Istanbul, Turkey*
ORCID ID: 0000-0001-8338-991X
** Corresponding Author Email: berna.altinel@marmara.edu.tr*

Machine Learning approaches. Our method also used Latent Dirichlet Allocation (LDA) [30, 31] which represents a document collection by their thematic topics. LDA sees usually a considerably large collection of documents as a mixture of topics. Each of these topics is shown as a distribution over the terms of a vocabulary. Based on this, using LDA we can determine a number of topics in a large document collection in an unsupervised way and assign one or more topics to each document.

2.1 Diffusion Process-Based Approaches:

In diffusion-based approaches, an attempt is made to understand the structure of the network and analyze how information is spread by simulating social networks. The influence maximization problem was firstly mentioned in [1] . This study rapidly attracted the attention of researchers studying in the field of opinion leader detection. For instance, the study [2] proposed a method called IM-LPA by combining the influence maximization algorithm with label propagation to rank the opinion leaders. In the study [3] , the role of opinion leaders in the diffusion of a product is researched. An empirical survey was made and as a result, three characteristics of opinion leaders were revealed. Then, a 3-step agent-based simulation model was constructed in which hypotheses would be tested. These steps are mass media, VoM (Word of Mouth), and adoption. As a consequence of the study, it was observed that the adoption speed of the product increased according to the presence of opinion leaders in the network and the appropriate comments of the opinion leader about the product. The basic difference between these studies and our study is the diffusion process since OLED does not contain any diffusion process in its methodology.

2.2 Graph-Based Approaches:

In graph-based approaches, network graphs are created based on the relationships between users, and opinion leaders try to be identified using user features and centrality measurements inferred from the graph. In the study [4], political opinion leaders were attempted to be detected by using the degree centrality, eigenvector centrality, and betweenness centrality of user nodes in the social network structure. The sample dataset created with 6000 users was increased to 15 million using two-degrees of separation. The users in the data set were filtered considering attributes such as language and location and there are 10 million users left. Top 100 users were listed using three centrality measurements and these measurements positively correlated. In another study [5], a method was developed that offers a probabilistic generate-graph model using user features and outbreak nodes instead of static features such as a number of good friends. Users were categorized using user attributes and it was found that efficient nodes had higher value by calculating outbreak index values [5]. The method in the study performed better when compared to Support Vector Machine (SVM) and Bayesian algorithm results.

In a very recent study [25] the authors identify important posts in social network community by taking weighted average of share, the number of comments and likes of each post since they give different importance to each post type. This feature is named as “ScoreImp”. Then they create the list of French opinion term list. On each post, they count these terms in text by string matching. This feature is named as “ScoreOp”. Then they calculate these features for each user. In order to determine influencer score (ScoreInf) the sum of ScoreImp and ScoreOp values are used in[25]. Three datasets are generated during three months from Facebook. Each dataset has nearly 600 posts and 1900 users. For evaluation, they calculate the number of important posts and the number of unimportant posts using ScoreInf for selected OLs. They compare their algorithm DDOL(Dynamic Detection Opinion Leader) with pagerank and betweenness centrality using precision. DDOL get better precision compared to betweenness centrality but DDOL performs slightly less precision than PageRank. Compared to pagerank, since DDOL is interested only with nodes and not with arcs, it is suitable to all graph types (Complex and Simple) and dynamic changes. In addition, it has less computation complexity compared to PageRank and Betweenness centrality.

There are some important differences between this study and our study. One of them is the authors in [25] do not aim to detect topic based opinion leaders since their dataset has only single main topic. Other difference is that [25] is not interested with graph arcs, user relations in graph. For future work, the authors want to add sentiment analysis or image analysis into their proposed method.

2.3 Statistical and Stochastic Approaches:

In statistical and stochastic approaches, various calculations and features are used to discover dependencies within networks and use them in opinion leader detection. In the work [6], a framework called BARR was proposed which determines opinion leaders and gives marketers a chance to determine their strategies according to the posts of bloggers. Blogs were searched using the user-defined keywords and web pages were analyzed. Domain ontology was extracted by calculating entropy values from the collected ontologies. Relations between bloggers have been found using domain name ontology, centrality, and prestige. Technique for Order Preference by Similarity to Ideal Solution that summarizes the Euclidean distance between measurements and ideal solution is used for hot blog selection by topics. Whether a user is an opinion leader has been decided based on the quantity and quality values calculated. The study [7] investigates which opinion leader is best for a selected market in terms of diffusion speed and the maximum cumulative number of adopters. Based on social network theory, it has been examined how opinion leaders affect the network and product diffusion. A simulation with three different scenarios was repeated 100 times with the network with 10000 entities.

The aim of another study [24] is to identify the group of users having the maximum synergy declared as the coalition of opinion leader. For this purpose, game theory approach is used. In methodology, they hypothesize that each user behaves like a player in the network in which trust and other centrality measures helps to find the marginal contribution of a user in the game. They propose an inventive and distinctive solution to measure the individual payoff using the distance-based centrality parameter. They also compute the Shapley value for each user to identify the maximum marginal contribution. They compare their results with other SNA measures (PageRank, Betweenness Centrality, Degree centrality, Eigenvector centrality, and Closeness centrality). They use two real networks: Wiki-vote and Bitcoin OTC trust weighted signed dataset. Wiki-vote dataset has 7115 nodes and 103 689 edges and Bitcoin dataset has 5881 nodes and 35 593 links. Their study provides superior results in terms of the accuracy, precision, time complexity, rate of convergence, and computational time with other SNA (Social Network Analysis) measures. One of the drawbacks of this work is that the proposed method is only suitable for the static network. For future work, they would take the challenge to overcome this issue. One of the advantages of this work is that they use degree of trust as the main component to gain maximum marginal contribution, which means that their proposed prototype represents real world scenario. There are two basic differences between this work and our works: 1.) they only examine social network graph, 2.) they do not use sentiment analysis and topic-based analysis.

In the work [26] the role of opinion leaders during COVID-19 outbreak in China are analyzed. They examine the public figures and official governmental accounts of Weibo microblogging platform as a case study. They conduct statistical analysis of important topics related to pandemic with literature references by focusing on effects of positive sentiments on the successful battle against COVID-19. Since it is not a technical paper, they do not implement any experiments and no datasets are collected.

2.4 PageRank Based Approaches:

In PageRank based approaches, researchers prefer to try to improve by making changes in the PageRank algorithm or to use it as the baseline for the algorithms they have developed themselves. In recent work [8] , and improved weighted LeaderRank algorithm has been represented. User weights were calculated using not only replies but also posting, reading, be praised, etc. The link between the users is based on whether one user is replying to another. After calculating the influence score of each user, it was compared with the influence values in PageRank and LeaderRank algorithms, and results that are more accurate were obtained with weighted LeaderRank. In the study [9], a novel algorithm named InfluenceRank, which ranks the blogs in the blog network according to their importance and correctness, has been studied. The topic space was created by accepting each entry in the blogs as a document and using LDA. Then, the feature vector of each entry was created in the topic space and dissimilarities were calculated using cosine similarity. The proposed algorithm performed better than PageRank, Random Sampling, Time-based Ranking, and Information Novelty-based Ranking algorithms used as the baseline.

In a recent study [23], rank after clustering (RaCRank) algorithm is proposed to detect opinion leaders in social networks. Algorithm consists of two phases, in the first phase, modified version of K-means is utilized with the following features: in degree, betweenness, center. Then, they proposed two-hop clustering coefficient. In the second phase of the algorithm, users' leadership score is calculated based on user activeness, user influence and center. Experiments are conducted by using social network with 49,613 users, and 59957 edges among these users. Suggested method is compared with AllUserRank, ClusterRank and UI-LR. Although RaCRank algorithm performs slightly worse than UI_LR, it outperforms AllClusterRank. According to the authors in [52], future studies can focus on detection of topic-based opinion leaders.

2.5 Machine Learning Approaches:

Machine learning approaches are used in works like classifying texts in the social network to try to detect opinion leaders. In a work [10] , the authors focus on how news spreads on Twitter. Two different SVM classifiers with the bag-of-words(BoW) method were trained for determining whether the news that shared on Twitter before the mass media was a rumor or not. While the first classifier determines the tweet's relevance to the topic, the second classifier decides the tweet is certain or uncertain. It was discovered that people prefer a small group of people called opinion leaders who share information with people on Twitter instead of learning from news sources. In another work [11] , OLFinder has been proposed to find influential users by analyzing important topics in the domain. Popularity score according to users' links on the network and competency score based on topics were calculated. According to the results obtained, the proposed method has outperformed the basic algorithms in the literature. LDA was used for topic extraction and was found to be better than TF-IDF.

In a recent work [22], a novel approach is proposed for community detection and social network-based nature-inspired whale optimization algorithm. Global and local top-N opinion leaders are detected. The community-partitioning algorithm is used to discover communities on the social network. The experiments are performed using two different data sets: The first one is synthesized data set consisting of 100 nodes and 467 edges, and the second is called 'wiki-vote data set' consisting of 7115 nodes and 103,689 edges. As the number of users on the network increases, the performance of the algorithm has increased.

In another study [27], an atomized three-layered platform is proposed to extract opinion leaders with the pre-defined topics. The layers in the suggested work are: 1) Collection layer for data mining and cleaning processes, 2) Classification layer for topic extraction and sentiment analysis, 3) Reasoning Layer for user inclination and social interaction analysis. Ground truth-based evaluation, survey-based evaluation and use-case scenarios evaluate the experimental results. The pipeline they proposed is similar to our work in overall, but there are some major differences in details. In general, opinion leader researches suffer from absence of labeled data for microblogging platforms. Therefore, they start with unsupervised clustering algorithm k-means + TF-IDF to extract topics in tweet data. To improve the accuracy of clustering, they enrich tweets they use Wikipedia articles. Our choices of algorithms are different in this step; we use LDA with supervision human experts for topic labeling and we use SVM-based classifiers for classification. For user inclination analysis, they used SVM with Bag-of-Words features to detect user topic, then adjust it with weighted neighborhood relations for label relaxation purpose. In this step, we use completely different set of features for users (i.e., three centrality measures and 4 topic-related features such as focus rate, activeness, authenticity, follower/following ratio). Since their framework mostly based on unsupervised clustering, they do not need huge datasets for training and test purposes unlike us. To evaluate topic labeling, they use 400 manually labeled tweets and 500 tweets for survey analysis. For user inclination, they collect 28132 tweets by 853 users and for use-case scenarios, they collect 164714 tweets posted by 67294 users with a keyword "headphone". Their main contribution is creating a framework for OL detection and evaluation with minimal human interaction. As a future work, they plan to create a graphical web interface for end-users with rich visualization.

3. Methodology

We propose a method to identify topic-based opinion leaders based on user features and the network structure. We specifically work on Twitter network, as it is the most convenient social network to collect public data and use follower and friend. Relations between users to build a social network graph fro our experiments. This system is called *Opinion Leaders Detection (OLED)*. Fig. 1 shows the basic steps of this approach. Four main steps will be explained in the following sections, respectively: Data Collection and Preprocessing, Topic Modeling, User Modeling and Detection of Topic-Based Opinion Leaders.

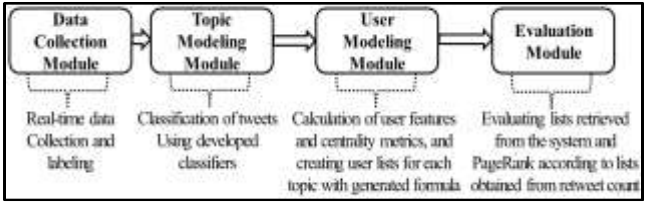

Data Collection Module
Topic Modeling Module
User Modeling Module
Evaluation Module
Real-time data Collection and labeling
Classification of tweets Using developed classifiers
Calculation of user features and centrality metrics, and creating user lists for each topic with generated formula
Evaluating lists retrieved from the system and PageRank according to lists obtained from retweet count

Fig. 1. Framework of OLED

3.1. Data Collection, Pre-processing and Clustering:

The OLED utilizes two different data sets. The first set consists of only tweets while the second set consists of both tweets and followers and friends' relations of the users. Both data sets are collected by using Twitter API. 350 Twitter users who are believed to be posting on various topics are manually selected to create these data sets. Next, 300 of the users are picked and their tweets are collected. At the end of this process, 1,842,499 tweets are collected. The remaining 50 users are used to build the second data set. Due to Twitter API's rate limits, 10% of the followers and friends of these users are collected. Next, the users who do not post in Turkish and whose profile is set to protected are removed. At this point, 5,924 users are collected. The previous steps are repeated by using this set of users to create the final version of the user network. As a result, 38,727 users and 97,842 edges among these users are collected. Next, tweets of these users are collected in real-time, between the dates December 1st, 2019 and January Dataset I only contains tweets of 300 users and it does not include relationship information of users. First dataset is used as training set to extract topic names by LDA and human-expert supervision. In order to avoid information leakage, we use a different set for our topic modeling and user modeling algorithms. In Dataset II, we need user-relationship information also so we collect friends and followers of 50 different users by randomly selected %10 edges due to twitter's API limitations. 5,924 and 32803 users are collected as first-degree and second-degree relations. Dataset II contains 38727 users and 97842 edges in total.31st, 2020. The resulting data set consists of 17,234,924 tweets.
The properties of datasets are shown in Table 1.

Table 1. Statistics About Datasets

	# users	# edges	# tweets
Data Set I	300	-	1,842,499
Data Set II	38,727	97,842	17,234,924

In the preprocessing step, mentions, hashtags, URLs, emojis, and punctuations are removed from the tweets. Additionally, lemmatization and stopword filtering are applied to the tweets. For lemmatization Zemberek NLP library [28] is used. After cleaning the data, the Latent Dirichlet Allocation tool called MAchine Learning LanguagE Toolkit (MALLET) [12] is used to determine the categories on the data set I.

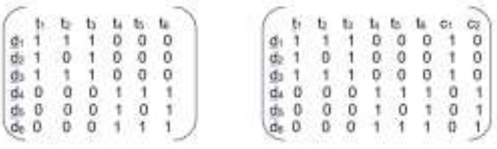
Table 2. Topic distribution of tweets after LDA results in Data Set I

Category	#tweets
Economy	300,335
Culture and Art	349,859
Politics	414,404
Sports	459,083
Technology	318,818

Since tweets are short text documents, one of the pooling methods proposed in [13] is applied to obtain more coherent clusters. According to this method, tweets are grouped based on each user, and each day. LDA is one of the most common topic modeling approaches. LDA uses a multinomial word distribution to represent topics semantically. The project team reviews word clusters generated by LDA and the clusters that can be logically labelled under a topic are extracted like it is applied in [13]. In other words, generated labels for data set I, is used for parameter optimization in classifiers in topic modelling part. These processes were repeated until a sufficient number of topics were obtained from LDA results. Lastly, the topic labels are added to the pooled tweets to construct the final data set. Five different topics are obtained: economy, culture and art, politics, sports, and technology. Tweet distribution based on these topics is shown in Table 2.

3.2. Topic-Based Tweet Classification:

In order to classify tweets according to their topic, several semantic kernel classifiers are applied. Two of them are Semantic Meaning Classifier [14] and Abstract Feature Classifier [15]. These semantic kernel classifiers attempt to add semantic values of terms into the classification process. Furthermore, Sprinkling and Adaptive Sprinkling versions of these semantic classifiers are also developed and conducted to the experimental environment.
Sprinkling (S) Method: In this method, the class label is added as a feature on the word-document matrix.



$$\left(\begin{array}{c|cccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ d_1 & 1 & 1 & 1 & 0 & 0 & 0 \\ d_2 & 1 & 0 & 1 & 0 & 0 & 0 \\ d_3 & 1 & 1 & 1 & 0 & 0 & 0 \\ d_4 & 0 & 0 & 0 & 1 & 1 & 1 \\ d_5 & 0 & 0 & 0 & 1 & 0 & 1 \\ d_6 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right)$$

(a)

$$\left(\begin{array}{c|cccccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & c_1 & c_2 \\ d_1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ d_2 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ d_3 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ d_4 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ d_5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ d_6 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{array}\right)$$

(b)

Fig. 2. (a) Standard word-document matrix, (b) Word-document matrix with class labels [16]

In Fig. 2, it is seen that the class label information is added to the word-document matrix as two separate properties. Since the first 3 documents belong to the c_1 class, they take the value 1 for c_1 and 0 for c_2 and the other 3 documents take the value 0 for the c_1 and 1 for the c_2 because they belong to the c_2 class [16]. In our study, separate features are added as class-labels into word-document matrix for each topic listed in in Table 2 (i.e., Economy, Culture and Art, Politics, Sports and Technology).
Adaptive Sprinkling (AS) Method: In real-world problems, it is often observed that some classes are more difficult to separate than others. In this method, the amount of sprinkling (the number of new columns to be added) is added as a parameter, which is directly proportional to the separability relationship of the respective class with the other classes [17] . Classifiers such as SVM produce a confusion matrix. Values not on the diagonal in this matrix indicate a difficulty in separating classes by the classifier. The number of sprinkling terms specific to the classes is produced by probabilistic calculations considering the confusion matrix in
Fig. 2. Next, Latent Semantic Indexing (LSI) is applied to the new data set. Before performing SVM on this new enhanced dataset, they drop columns that have already been added by the sprinkling method.
By using Sprinkling and Adaptive Sprinkling versions of these Semantic Meaning Classifier [14], Abstract Feature Classifier [15] classifiers, experiments are done using Dataset I. After that, according to the experimental results on Dataset I, the best performing classifier is decided and it is used in order to classify unlabeled instances in Dataset II.

3.3. User Modelling:

Constructing the Feature Set of Each User:

After labeling tweets on the topic-modeling step, subnetworks are constructed based on the topic distribution of tweets of each user. If the tweet percentage of a user for any topic exceeds the predefined threshold, the user is added to the subnetwork of that topic. The distribution of users are as follows: 692 users for Economy category, 3529 users Culture and Art category, 25193 users Politics category, 3067 users Sports category and 452 users Technology category.

Centrality measures of the users are calculated using these subnetworks. These centrality measures are used to indicate a user's location in the network and the degree of importance in the information flow. "Degree Centrality (dc)" indicates the number of nodes to which the node on a network is in a direct relationship. "Betweenness Centrality (bc)" determines the importance of a node that connect different nodes thanks to their location on a network. People with a high value of betweenness centrality are critical people because of their control over the information passed among others. The direct and indirect connections of the users with high "Closeness Centrality (cc)" value on a network allow them to reach other users more quickly. The more central a node is, the closer it will be to all other nodes.

Centrality values of the users of an example small subset from our network are shown in Table 3:

Table 3. A small set of users and their centrality values.

Username	Degree Centrality	Closeness Centrality	Betweenness Centrality
$User_1$	**286**	**0.397679**	**475853**
$User_2$	279	0.393008	458444
$User_3$	233	0.362291	441858
$User_4$	240	0.354847	400016
$User_5$	262	0.391786	374967
$User_6$	262	0.37557	366323

According to Table 3, $User_1$ has the highest degree centrality, this shows that $User_1$ has higher number of connections than others. $User_1$ also has the highest betweenness centrality, so much of the all communication pass through him. And finally, the highest closeness centrality also belongs him so he has easy access to all other nodes. Thus, we can say that User1 is a proper candidate for OLED in this small network if we only consider network features.

After the calculation of centralities, several user features are also included to feature set. The focus rate is used to calculate the topic distribution of the user's tweets [18] . For each topic, the user's focus rate (fr^t_u) is calculated by dividing the number of tweets posted for the topic (p^t_u) by the total number of tweets (p_u). If it exceeds the predefined threshold in any topic, it is assumed that the user is focused on this topic.

$$fr_u^t = \frac{|p_u^t|}{|p_u|} \tag{1}$$

Activeness is used to calculate how often a user tweets on a topic [18] . For each topic, the user's activeness (ac^t_u) is calculated by dividing the number of days posted for the topic (d^t_u) by the total number of days (d).

$$ac_u^t = \frac{|d_u^t|}{|d|} \tag{2}$$

Authenticity(au^t_u) is used to measure the originality of tweets about a topic [18] . User's retweets (rt^t_u) on a topic are subtracted from all tweets related to that topic (p^t_u), and then the result is divided by all tweets about that topic (p^t_u).

$$au_u^t = \frac{|p_u^t| - |rt_u^t|}{|p_u^t|} \tag{3}$$

The follower/following ratio is used to compare the number of users following a user with the number of users followed by that user [19] . For each topic, the user's follower/following ratio (ff^t_u) is calculated by dividing the user's indegree for the topic (id^t_u) by out degree (od^t_u).

$$ff_u^t = \frac{|id_u^t|}{|od_u^t|} \tag{4}$$

3.4. Detection of Topic-Based Opinion Leaders

K-Means Clustering algorithm is employed after a feature vector is created for each user by considering the subnetwork that a user is in. Opinion leaders show similar behaviors. For example, their focus rate and authenticity values are high because opinion leaders prefer to focus on a topic and to express their own opinions. Hence, using the clustering algorithm, it is attempted to find user groups that are more likely to be opinion leaders. In order to decide which clusters to choose as candidates, fuzzy-based methods are used as applied in [20] .

$$\boldsymbol{f(x) = \int_0^{x_{max}} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx} \tag{5}$$

A user must have high values in the feature vector to be considered an opinion leader. Therefore, normal cumulative distribution in Eq. (5) is used for each feature as a fuzzy membership function. In Eq. (5), μ shows the mean of the feature, and σ is the variance. x_{max} is the maximum value of the feature in the data set and x is the corresponding attribute. Then, a score is calculated for each cluster by multiplying the function result of each attribute in the feature vector. The clusters are sorted based on their scores and possible clusters are selected as candidate opinion leaders. Eq. (6) is generated to find actual opinion leaders.

$$\boldsymbol{OLscore_u^t = (w_1 \times dc) + (w_2 \times bc) + (w_3 \times cc) + (w_4 \times fr_u^t) + (w_5 \times ac_u^t) + (w_6 \times au_u^t) + (w_7 \times ff_u^t)} \tag{6}$$

where *w* vector can be interpreted as feature importance and gives the weights of features in our OLED formula which is just a weighted average of these features.

Users in possible clusters are labeled as OL (opinion leader), others as normal users, and feature weights obtained by training SVM are considered as coefficients in Eq. (6). User lists are ranked after each user's opinion leader score is calculated. Finally, the top number of users are selected as the opinion leader.

4. Experiments

3.1. Experiment Setting and Evaluation:

In pre-processing part, we run MALLET LDA multiple times with different parameter values in order to determine final categories in data set I. For alpha parameter 50 and 200 are used. 10 is used for optimization interval parameter which optimizes hyper parameters each iteration. For the number of topic parameter 10,15 and 20 values are used and number of word parameter selected as 20.

Due to the network size of the Data set II, the user-modeling module is implemented by utilizing the Apache Spark ecosystem. Neo4J is used as data storage and centrality metrics are calculated by using its query language Cypher. As a result, the feature vector for each user-topic pair is created.

In order to evaluate the lists of the opinion leaders retweet count is used. The total retweet count of a user is the number of times that the user retweeted by other users for the given topic.

$$\boldsymbol{RTcount^t(u) = \sum_{\substack{tweet\, \epsilon \\ (|p_u^t| - |rt_u^t|)}} + \sum_{\substack{u'\epsilon \\ subnet.}} rt_{u'}^{tweet}} \tag{7}$$

In Eq. (7), 'tweet' in the first sigma symbol indicates a user's own tweets, and the u' in the second sigma symbol is used for other users in the topic subnetwork. $rt_{u'}^{tweet}$ indicates if this tweet was retweeted by user u' and takes a value of 0 or 1. As a result of the calculation, the number of times the user's tweets are retweeted in the subnetwork is found. Then users are sorted by this measure in descending order and top N of them are selected. In the same manner, top N users are selected according to their score, which is calculated by the proposed method. Next, the agreement between the two lists of users is calculated.

3.2. Experimental results and discussion:

PageRank is chosen as a baseline algorithm to compare the performance of the OLED. The main functionality of the PageRank algorithm is to rank websites based on the network structure. It basically evaluates the value of a web page based on the quantity and quality of hyperlinks coming from other pages [21]. PageRank is a recursive function, which is defined by Google [21] as follows:

$$PR(A) = (1\text{-}d) + d \; (PR(T_1) / C(T_1) + ... + PR(T_n) / C(T_n)) \tag{8}$$

where Page A represents a page on web, T1 … Tn show the number of incoming citations to page A, C(A) represents the number of out-going citations from page A. d is a dumping factor and its default value is 0.85 is used in our work as a baseline to compare OLED algorithm.

The algorithm can be applied to any network. In literature, it is also used as a baseline in studies such as [18] where opinion leaders are determined by using a social network structure. Users on each topic are sorted based on their PageRank scores and then top number of users are selected as opinion leaders.

Table 4. Similarity score (%) of PageRank and OLED algorithms in terms of ReTweet count results

Method	**Top Number of Users**								**Category**
	30	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
PageRank	30	35	40	40	44	44	48	48	**Economy**
OLED	40	43	42	43	50	53	58	54	
PageRank	07	08	06	03	07	08	09	11	**Culture and Art**
OLED	07	10	14	12	17	23	24	27	
PageRank	30	30	30	26	24	24	23	21	**Politics**
OLED	33	38	32	33	31	30	31	28	
PageRank	13	15	16	15	19	19	23	26	**Sports**
OLED	20	20	22	27	30	29	29	29	
PageRank	30	30	34	40	43	48	48	48	**Technology**
OLED	30	35	36	43	49	49	49	49	
PageRank	***22***	***37***	***25***	***24***	***27***	***28***	***30***	***30***	***Average***
OLED	***26***	***42***	***29***	***31***	***35***	***30***	***38***	***37***	

The results obtained with ReTweet(RT) count, which is the first method, are shown in Table 4. As can be seen in Table 4, OLED exceeds or gives comparable results to the PageRank in almost all top N values in all of the topics. Almost in all N values and all topics OLED performs better. For N=40, there are some topics like Culture&Art, Economy and Politics in which OLED stays behind of baseline. We can explain it by the nature of Turkish Twitter ecosystem. People's interest are more likely in politics and sports and that's why we have low number of users in Economy (692), Culture&Art (3529) comparing to them in our Dataset II. It may cause our algorithm to perform poorly since it has semantic layers in it while Pagerank is a pure network-oriented method. The results increase as the value of N rises as the number of people considered opinion leaders in the lists is also approaching the number of users present in the network. Another reason is that, since the users are ranked according to the values they have in the algorithm, their order in the lists can be different and as the N increases, these different ranks are better captured.

We also compare our study against an existing study in the literature. In that study [32], a learning system was generated for the RepLab 2014 author profiling task at UNED. The features used for this system are tweet texts' POS tags, number of mentions, number of hashtags, number of links, number of emoticons, number of followers and retweet speed. Because we already have similar features in our dataset presented in this study, we tried to conduct the algorithms (i.e., random forest algorithm and naive Bayes algorithm from WEKA) presented by [32] in our experimental environment. We applied these algorithms to our *Politics* dataset, implementing the algorithms in Python with the scikit-learning library. The experimental results show that, the precision values of the random forest algorithm, naive Bayes algorithm, and OLED are 0.76, 0.63, and 0.80 with 100 opinion leaders.

5. Concluding Remarks

We propose a novel methodology for topic-based opinion leader detection with topic modeling and user modeling. In topic modeling part, we classify tweets with an advanced system of classifiers using several semantic kernels, and their Sprinkling and Adaptive Sprinkling versions. In user-modeling part, we construct a feature set for each user in the social network, which is built from the collected tweets such as focus rate, activeness, authenticity, follower/following ratio. These features also include network centrality metrics from social network analysis domain such as Degree Centrality, Betweenness Centrality, Closeness Centrality.

We perform our experiments on a data collection gathered from Twitter that includes 17,234,924 tweets and 38,727 unique users. According to topic modeling and user modeling results, we give leadership scores to each user in the network. Users with highest scores are stated as opinion leaders. In order to evaluate OLED's performance, we also run PageRank algorithm on the same dataset. Comparison of the results obtained from PageRank and OLED is a very tough job since there is no existing standard technique to evaluate opinion leaders. In the literature, different approaches are used in many altered studies because this evaluation can be subjective and specific to the problem domain. Therefore, the evaluation of the results in this study is done retweet count to prove that OLED outperforms PageRank.

According to experimental results, our framework OLED shows remarkable performance in compare to PageRank in majority of topics and all selected top number of opinion leaders. OLED outperforms PageRank in all categories in our dataset. These categories are *Economy, Culture-Art, Politics, Sports and Technology* as shown in Table 4. We also report the average scores for each top number of users for each category. For example, the average scores of OLED and PageRank with 60 top number of users are 24 and 31; respectively.

These preliminary results motivate us to improve our model with the contribution of some other user features especially in user-modeling part. Furthermore, creating a system, which has the capability of finding real-time opinion leaders in a dynamic network, might also be a good future work item.

Acknowledgment

This work is supported in part by The Scientific and Technological Research Council of Turkey (TÜBİTAK) grant number ARDEB-118E315.

References

[1] D. Kempe, J. Kleinberg and É. Tardos, Maximizing the spread of influence through a social network. In Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining, 2003, pp.137–146.

[2] Y. Zhao, S. Li and F. Jin, Identification of influential nodes in social net- works with community structure based on label propagation. Neurocomputing, 2016, 210, pp.34–44.

[3] V. Eck, P. S., W. Jager and P. S. Leeflang, Opinion leaders' role in innovation diffusion: A simulation study. Journal of Product Innovation Management, 2011, 28(2), pp.187-203.

[4] O. Z. Gökçe, E. Hatipoğlu, G. Göktürk, B. Luetgert and Y. Saygin, Twitter and politics: Identifying Turkish opinion leaders in new social media. Turkish Studies, 2014, 15(4), pp.671-688.

[5] L. Cui and D. Pi, Identification of Micro-blog Opinion Leaders based on User Features and Outbreak Nodes. International Journal of Emerging Technologies in Learning, 2017, 12(1).

[6] F. Li and T. C. Du, Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. Decision support systems, 2011, 51(1), pp. 190-197.

[7] Y. Cho, J. Hwang and D. Lee, "Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach." Technological Forecasting and Social Change, 2012, 79(1), pp.97-106.

[8] L. Luo, Y. Yang, Z. Chen and Y. Wei, Identifying opinion leaders with improved weighted LeaderRank in online learning communities. International Journal of Performability Engineering, 2018, 14(2), pp.193-201.

[9] X. Song, Y. Chi, K. Hino and B. Tseng, Identifying opinion leaders in the blogosphere. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007, pp. 971-974.

[10] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko and K. L. Ma, Breaking news on twitter. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, pp. 2751-2754.

[11] A. Aleahmad, P. Karisani, M. Rahgozar and F. Oroumchian, OLFinder: Finding opinion leaders in online social networks. Journal of Information Science, 2016, 42(5), pp. 659-674.

[12] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu, 2002.

[13] Z.Z. Alp, S.G. Ögüdücü, Extracting topical information of tweets using hashtags, in: 14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9–11, 2015, 2015, pp. 644–648, doi: 10. 1109/ICMLA.2015.73.

[14] B. Altınel, M.C. Ganiz and B. Diri, "A Corpus-Based Semantic Kernel for Text Classification by using Meaning Values of Terms", Elsevier, Engineering Applications of Artificial Intelligence Volume 43, August 2015, pp. 54–66.

[15] B. Altınel, B. Diri, M.C. Ganiz, "A Novel Semantic Smoothing Kernel for Text Classification with Class-based Weighting". Knowledge-Based Systems, 2015, Vol. 89, pp. 265-177.

[16] S. Chakraborti, R. Lothian, N. Wiratunga and S. Watt, "Sprinkling: supervised latent semantic indexing". In European Conference on Information Retrieval, 2006, pp. 510-514, Springer, Berlin, Heidelberg.

[17] S. Chakraborti, R. Mukras, R. Lothian, N. Wiratunga, S. N. Watt and D. J. Harper, "Supervised Latent Semantic Indexing Using Adaptive Sprinkling", 2007, In IJCAI, Vol. 7, pp. 1582-1587.

[18] Z.Z. Alp, S.G. Ögüdücü, "Influence factorization for identifying authorities in twitter". Knowledge-Based Systems, 2019, 163, 944-954.

[19] I. Anger and C. Kittl, Measuring influence on Twitter. In Proceedings of the 11th international conference on knowledge management and knowledge technologies, 2011, pp. 1-4.

[20] J. Duan, J. Zeng and B. Luo, Identification of opinion leaders based on user clustering and sentiment analysis. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014, Vol. 1., pp. 1-5.

[21] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: Bringing order to the web", 1999, Stanford InfoLab.

[22] L. Jain, R. Katarya & S. Sachdeva, Opinion leader detection using whale optimization algorithm in online social network, Expert Systems with Applications, 142, 113016, 2020.

[23] B. Zhang, Y. Bai, Q. Zhang, J. Lian & M. Li, An Opinion-Leader Mining Method in Social Networks with a Phased-Clustering Perspective, IEEE Access, 8, 31539-31550, 2020.

[24] L. Jain, R. Katarya & S. Sachdeva, Recognition of opinion leaders coalitions in online social network using game theory. Knowledge-Based Systems, 2020, 203, 106158.

[25] W. Oueslati, S. Arrami, Z. Dhouioui & M. Massaabi, Opinion leaders' detection in dynamic social networks. Concurrency and Computation: Practice and Experience, 2020, e5692.

[26] W. Fang, B. Gao & N. Li, Analysis of the Influence of Opinion Leaders on Public Emergencies through Microblogging. *Open Journal of Social Sciences*, 2020, *8*(5), 154-158.

[27] C. L. Yang, Novel platform for topic group mining, crowd opinion analysis and opinion leader identification in on-line social network platforms, 2020.

[28] A. A. Akın & M. D. Akın, Zemberek, an open source nlp framework for turkic languages. Structure, 2007, 10, 1-5.

[29] A. Müller, Referral marketing on social media platforms—guidelines on how businesses can identify and successfully integrate opinion leaders in their online marketing strategy. In Omnichannel Branding, 2018, pp. 131-171, Springer Gabler, Wiesbaden.

[30] D. M. Blei, A. Y. Ng & M. I. Jordan, Latent dirichlet allocation. Journal of machine Learning research, 2003, 3(Jan), 993-1022.

[31] D. M. Blei, Probabilistic topic models. Communications of the ACM, 2012, 55(4), 77-84.

[32] J. Jesus, M. Lomena and L. Ostenero, F. UNED at CLEF RepLab: Author Profiling, in Proceedings of the Fifth International Conference of the CLEF initiative. Springer, 2014, pp. 1537-1549.