A Hybrid Credibility Analysis Method Applied on Turkish Tweets with TV News and Discussion Programs Related Content

Ali Fatih Gündüz[1] and Pınar Karagöz[2]

[1]Akçadağ Vocational School, İnönü University
[2]Computer Engineering Department, Middle East Technical University
[1]Akçadağ, Malatya, Turkey
[2]Çankaya, Ankara, Turkey
fatih.gunduz@inonu.edu.tr, karagoz@ceng.metu.edu.tr

ABSTRACT

In this paper, credibility analysis of microblog messages is studied. We collected our data from one of the most important microblogging services, Twitter. Our data set is created from the Turkish tweets written for weekly television programs broadcast in Turkey with political, cultural and/or financial contents. We adapted a new credibility definition based on three dimensions: being free from offensive words, not being spam and being newsworthy. To analyse credibility of tweets, we proposed a method by hybridizing supervised learning based techniques with graph based techniques. Moreover we used content based and collaborative filtering based techniques in this hybrid method. The proposed method consists of two phases: supervised learning phase and graph based improvement phase. Supervised classification algorithms are applied in the first phase and the results are improved in the graph based phase of the method. We focused on tweet-tweet, tweet-writer and writer-writer relations in the second phase of our study. The performance of the proposed method is measured by comparing it with human volunteers' overall evaluation at the end.

KEYWORDS

Twitter, social media, information, credibility, microblogs, data mining, machine learning, natural language processing, collaborative filtering, cosine similarity, classification, content based filtering

1 INTRODUCTION

Internet based communication tools provides a perfect research area and information credibility is an important aspect of communication. From past to today, we observed change in the technology and the tools of communication. However measuring believability of messages always remained as an important and crucial task. Today internet based social media tools are on the rise not only in Turkey but also in the world. The number and variety of web-based media platforms available is impressive.

Social media became extremely popular for many reasons. Following personal and professional pursuits, building and sustaining friendship relations, advertising business promotions and many other purposes are met by those online microblogging services. Another reason as to why social networking attracted public attention is the ability to post about spontaneous events easily. People can express their feelings by sharing messages with many topics ranging from their daily and ordinary activities to cultural and social events.

Twitter[1] is one of the most important microblogging services and it allows its users to share 140 character long messages called tweets. Twitter has been providing an online environment in which people can share their ideas, comments and concerns since 2006. Today it is a huge company with 313 million active users. Every day millions of tweets are written in this online environment about many topics.

Users of Twitter can read, favourite and share tweets of other users. Moreover users can create friendship networks among themselves by following each other. Those friends can enjoy

[1] https://about.twitter.com/tr/company

creating private or public messages. To send public messages, users mention about their friends' user names in tweet content by adding '@' character in front of it. This practice is called as mentioning and this tag is called as mention tag. Similarly users can use another tag by adding '#' character in front of a specific word to create hashtags. Twitter displays the tweets containing same hashtags together so that users can follow and contribute to specific discussions.

Twitter enables researchers to read, query and collect tweets of users who do not disable public visibility of their statuses. Our data set is created from those publicly written tweets which are tweeted for news and discussion programs broadcast on the television weekly. Many TV programs ask and encourage their audience for participation through Twitter accounts of programs. Most of them have Twitter accounts in order to enable their audience to contribute in the program flow by asking questions, making comments and expressing their feelings by writing tweets consisting of program specific hashtags and/or mention tags. Hosts of those programs read those tweets and direct the program flow accordingly if they desire to do so. However separating junk from useful information in the tweet flood is a big challenge and time consuming.

To analyse the credibility of a tweet, we proposed a hybrid method in this study. Content based techniques and collaborative filtering based techniques are hybridized in our method. Moreover, we improved our method by deploying word-checking algorithms from a slang-word dictionary. The proposed method is composed of two phases namely supervised learning phase and graph based improvement phase. Firstly, in supervised learning phase, we applied machine learning classification algorithms on the tweet set and then the classification results are improved in the graph based phase. We examined both tweets and users in order to create a connected graph from them with links between user-user, user-tweet and tweet-tweet elements. This graph is used to investigate whether credible tweets are contextually similar and whether writers of credible tweets form a closer clique in the graph. In order to achieve this, we created links between tweet-tweet nodes only if their contents' cosine similarity is bigger than a predefined threshold. Also we used followership relations between authors to create user-user links between nodes. Finally tweet-author relations are used to link a tweet node with a user node in the graph.

On the other hand, we proposed a definition for credibility as well. Credible is defined as "able to be trusted or believed" by Cambridge dictionary[2]. By its nature, credibility is a subjective matter and it is always open to discussion. In addition, its measurement depends on individual opinions and changes greatly from person to person. Fogg and Tseng [1] state that credible information is believable information and they describe credibility as a perceived quality composed of multiple dimensions. So we based credibility definition on three dimensions: being free from offensive words, being free from spamming and being newsworthy.

Our study is conducted with binary rating while deciding dimensional values of credibility. The proposed hybrid method is used to assign yes or no statuses to each dimension. Each one of those three dimensions is examined separately and final credibility decision is made according to those three dimensions' overall results. A tweet is labelled as credible only if it is free from offensive words, free from spams and it is newsworthy. The performance of the proposed method is analysed by comparing it with the human volunteers' classification results. Since those dimensions are likely to be interpreted variously by different people, each tweet of our data set is read by three volunteers and majority voting is used for each dimension.

2 RELATED WORKS

In 2012, Kang et al. [2] proposed two definitions for tweet credibility as "degree of believability that can be assigned to a tweet about

[2] dictionary.cambridge.org/dictionary/english/credible

a target topic" and "expected believability imparted on a user as a result of their standing in the social network". In addition to these, they stated later that [3] credibility is a function of perception consisting of the object being perceived and the person who perceives it.

Fogg [4] expressed website credibility in terms of prominence and interpretation which are defined as likelihood of being noticed and judgement of the people who noticed the element in his study.

Castillo and Yamaguchi studied both credibility assessment and newsworthiness of tweets [5]. In their study, they focused on credibility of information and used the term credibility in the sense of believability. They classified tweets as credible or not. They randomly selected 383 topics from Twitter Monitor[3] [6] collection and get it evaluated by Mechanical Turk[4] by asking evaluators if they consider that a certain set of tweets as newsworthy or only informal conversations. Then they asked another group to read the text content and state if they believe that those tweets are likely to be true or false. In this evaluation, they considered four levels of credibility and asked evaluators to provide justification in that fuzzy format. They proposed a supervised learning based method to automatically assess the credibility level of tweets which has a precision and recall rate between 70% and 80%.

Detecting and preventing spam is another aspect of credibility. Not only individuals write those spam tweets but also designed tweet generator tools are used to carry out this annoying and potentially malicious activity. Ferrara et al. [7] stated that hundreds of thousands of social, economic and political incentives presented by highly crowded social media ecosystems attract spammers to design human imitating bot algorithms. Forelle et al. [8] stated that bots are used for political lobbying in several countries like Russia, Mexico, China, UK, US and Turkey.

Twitter attaches importance to the fight against the spammers in order to sustain a spam-free social environment. They encourage[5] their users to report both profiles and individual tweets for spamming. Moreover they present technical solutions such as link shortener (t.co) to detect whether links lead to malicious contents as well.

To detect Twitter spam, there are two different approaches in the literature: focusing on the user classification and examining tweet content. In the first approach, profile details of the user, number of followers and friends, recent activities in the previous weeks, user behaviours and tweeting frequencies are investigated. Studies like [9], [10] and [11] aimed to classify users as spammers and non-spammers according to these user attributes. The second approach considers topics of the tweets, duplications between the tweets, urls in the tweets, number of words and characters in the texts. Martinez et al. [12] presented an example of this approach in which they detected spam tweets without any previous user information but by using contextual features obtained by natural language processing. Clark et al. [13] proposed a solution to the problem of separating automated spam generators from human tweeters by a classification algorithm operating by using linguistic attributes like url count, average lexical dissimilarity and word introduction rate decay.

There are hybrid solutions of user based and content based approaches like [14] and [15]. Bara et al. [15] proposed a three step solution in which they firstly look for malicious links provided by Twitter database, secondly they look for pattern similarities between spam tweets and original tweets and finally they construct a bipartite network between users and corresponding tweets.

Pal et al. [16] studied tweet credibility from another perspective by classifying tweet writers. In their study, they tried to find most interesting and authoritative authors among millions of Twitter users for given specific topics. They computed self-similarity score for authors between their last two tweets so as to measure how similar an author writes. This score is

[3] http://www.twittermonitor.net

[4] http://www.mturk.com/

[5] https://support.twitter.com/articles/64986?lang=en

used to explore the width of the users' interest area. They also classified tweets into three categories: original tweets, conversational tweets and repeated tweets. They counted the number of tweets in different categories of authors while deciding about their interestingness for clustering the users.

Other than classification approaches, there are graph based solutions as well. Graph based solutions are basically use variations of well-known PageRank [17] and HITS [18] algorithms. Page and Brin, with PageRank, aimed to measure and rate relative importance of Web pages mechanically. In this algorithm, the linking design among the web pages is considered in a graph structure. Being query independent and more sophisticated than simply counting links, PageRank ranks pages according to their importance of back links and forward links which directs to and are directed from the web pages. With HITS algorithm, Kleinberg [18] aimed to extract information from the link structure of network environment too. Although HITS is not solely specific to WWW, aiming to improve web search systems it identifies two kinds of web pages: authorities which are the pages that users look for to reach information and hubs which are pointer pages that lead to authorities. Kleinberg focused on the mutual relationship between those two kinds by giving non-negative invariant weights to each node and then making iterative score transfers between interlinked hub and authorities until scores converge to the equilibrium values.

Another graph based study is TURank which constituted a base to our study. Yamaguchi et al. [19] proposed Twitter user ranking algorithm (TURank) to determine authoritative users. They defined authoritative users as the ones who frequently submit useful information and they aimed to measure authoritativeness of users in order to rank them. They constructed a user-tweet schema graph where nodes are created from users and tweets. On the other hand edges are created from post, posted by, follow, followed by, retweet and retweeted by relations between user-tweet, user-user and tweet-tweet nodes. Then they applied ObjectRank [20] on the user-tweet schema graph to evaluate the users' authority scores.
Similarly, Gun and Karagoz [21] proposed a hybrid solution combining feature based and graph based methods for credibility analysis problem in microblogs. They focused on message, user and topic relationship in the graph based part of their study. They gathered 43 feature attributes from tweet, topic and user data in order to use them in feature based classification. They tried to label tweets as newsworthy, important and correct for determining which information in Twitter is credible.

3 DATA COLLECTION AND CONSTRUCTION OF THE GOLD STANDARD

Our data set consists of tweets, their authors and the ground truth evaluations obtained from volunteers. In order to carry out this study, we crawled tweets with specific query keywords related with weekly Turkish television programs. We selected television programs with political, social, economic and cultural contents. Concepts of the selected programs is built upon discussions between experts who are hosted by the channel or presentations of celebrities about mentioned topics. Those programs are open to audience contributions through Twitter and the hosts read comments and direct questions to guests during the program flow if they desire to do so.
The crawled query keys are explicit hashtags and/or mention tags used by the program producers so that they receive comments and questions from their audience through Twitter. We only gathered the tweets which are deliberately written for the selected television programs. During data collection period we crawled tweet id, tweet text, user id, retweet and favorite counts of the tweet. Tweet text is parsed and 22 different features obtained about the tweet such as length of tweet, number of words in it, fraction of upper case letters, fraction of tagged words as hashtags and mention tags, whether tweet contains question mark, exclamation mark and whether emoticons exist in

tweet etc. Moreover positive and negative sentiment scores of the tweets are obtained from SentiStrength API[6] and added to the feature set.

Other than tweet, data of the users are crawled as well. Collected user features are friend count, follower count, tweet count, friend and follower lists[7]. The follower and friend lists are used to determine user-user links of the data's internal friendship network.

In order to construct the gold standard for the evaluation, we conducted a user study with contribution of volunteers. Each tweet is read by three people and they answered to the questions: [8]

1. Does the tweet contain swearing, abusing or offensive words?
2. Is the tweet written for distracting, unrelated, advertising or out of program scope purposes?
3. Is the content interesting, important or news-worthy?

The volunteers answered each of these questions as either Yes or No. The ground truth label is determined by using majority voting. Each question is experimented separately and formed a dimension of this study. Finally we labelled a tweet as credible only if both three dimensions of it provide proper answers. Tweets classified as 'no' with respect to the first two questions and 'yes' with respect to the last question are labelled as credible while the others are identified as ineligible.

4 METHODOLOGY

Tweet data collection and ground truth evaluations form the first chapter of our study. The next two chapters are the phases of the proposed method which are applied respectively. In the first phase we used supervised learning techniques and then the obtained results were improved in the graph based part. For the offensiveness analysis of the tweet we also experimented slang word dictionary based methods and compared performances of them.

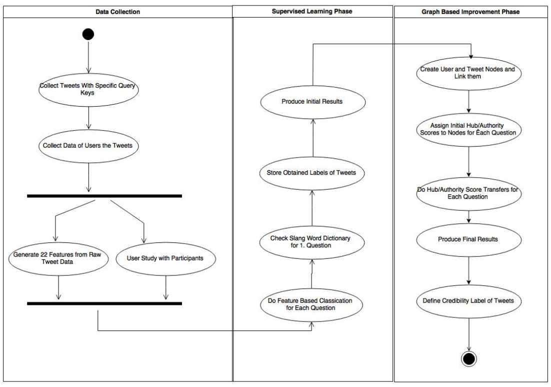
Data Collection
Supervised Learning Phase
Graph Based Improvement Phase
Collect Tweets With Specific Query Keys
Collect Data of Users the Tweets
Generate 22 Features from Raw Tweet Data
User Study with Participants
Do Feature Based Classication for Each Question
Check Slang Word Dictionary for 1. Question
Store Obtained Labels of Tweets
Produce Initial Results
Create User and Tweet Nodes and Link them
Assign Initial Hub/Authority Scores to Nodes for Each Question
Do Hub/Authority Score Transfers for Each Question
Produce Final Results
Define Credibility Label of Tweets

Figure 1. Activity diagram of the proposed credibility analysis system.

4.1 First Phase - Supervised Learning

In this study we used Weka[9] API for Naive Bayes, kstar, ADTree and J48 decision tree classification algorithms. We used 10 folds cross validation in this phase and obtained best results with J48 decision tree algorithm for both three dimensions. The features used in this phase is shown in Table 1.

4.2 Second Phase - Graph Based Improvement

After applying the supervised learning phase of the proposed method, aiming to improve classification results we applied second phase of our method on the data set. Firstly, converting tweets and users to nodes we created a connected undirected graph. Secondly, we assigned initial scores to those nodes according

[6] http://sentistrength.wlv.ac.uk/

[7] Friend refers to the users followed by the user and followers refers to the users following the user

[8] As our tweet data were constructed from Turkish tweets, the questions above were Turkish in our website and volunteers were native Turkish speakers. Original questions in Turkish were:

1. Küfür, Hakaret, Saldırgan veya İncitici İfade İçeriyor mu?
2. Dikkat Dağıtıcı, Alakasız, Reklam İçerikli veya Program Dışı Bir Amaçla mı Yazılmış?
3. İçerik İlginç, Dikkate Değer veya Haber Değeri Taşıyor mu?

[9] Weka. http://www.cs.waikato.ac.nz/ml/weka

Table 1. Features of tweets used at supervised learning phase

No	Feature
1	Length of tweet
2	Fraction of upper case letters
3	Total number of words
4	Number of words with mention tags
5	Number of words with hashtags
6	Number of words without '@'/'#' tags
7	Fraction of tagged words
8	Whether contains question mark
9	Whether contains exclamation mark
10	Whether contains smile emoticon
11	Whether contains frown emoticon
12	Whether contains URL
13	Positive sentiment score
14	Negative sentiment score
15	Whether contains first pronoun
16	Whether contains second pronoun
17	Whether contains demonstrative pronoun
18	Whether contains interrogative pronoun
19	Retweet count
20	Is retweet
21	Is reply to a user
22	Favorite count

to the results of the first phase and run random walk iterations on this graph.

4.2.1 Graph Construction

During graph construction, we linked nodes according to following rules:

1. A user node is directly linked to a tweet node if the user is the writer of the tweet.

2. A user node is directly linked to another user node if the other user exists in the follower/friend list of the user.

3. A tweet node is directly linked to another tweet node if the tweet's content has equal to or more than a predefined cosine similarity with the other tweet's content.

By linking users together we aimed to investigate whether credible tweets are written by similar people that are friends and followers of each other. On the other hand, we examined whether credible tweets are contextually similar or not by linking those tweets in the graph.
To this aim, we first parsed the text of the tweets and obtained the word sets and eliminated the effect of stop words. Those word sets were processed with Zemberek[10] Turkish NLP tool and we replaced them with their corresponding longest lemma term so that we could identify relations among the same words in different morphological forms. This textual data is converted to term vector for each tweet.
Term vector of a tweet contains longest lemmas of all unique words existing in its text and corresponding term frequency-inverse document frequency multiplication score pairs.
In order to obtain multiplication results firstly we calculated the term frequencies of the longest lemma terms of tweets according to Equation 1.

$$TF(T_i, w_j) = \frac{Number\ of\ times\ word\ w_j\ occurs\ in\ T_i}{Number\ of\ words\ in\ tweet\ T_i} \tag{1}$$

Then inverse document frequencies of words are calculated according to Equation 2.

$$IDF(w, D) = log_{10}(\frac{TNTV}{CTV}) \tag{2}$$

where TNTV is total number of term vectors in dataset D and CTV is the number of term vectors which contain the word w.
Those terms and their corresponding term frequency-inverse document frequency multiplication result pairs are used to obtain Tf-Idf based term vectors of tweets according to Equation 3.

$$TfIdf\ Based\ Term\ Vector\ of\ T_i = \Big((w_1, tfidf_1), (w_2, tfidf_2), ..., (w_n, tfidf_n)\Big) \tag{3}$$

Finally, for each tweet, we calculated cosine similarity of its term-vector with all others according to Equation 4. Depending on the

[10] Zemberek Project, https://github.com ahmetaa/zemberek-nlp

cosine similarity value, we linked associated tweet nodes in the graph.

$$Cosine\ similarity(Tweet_i, Tweet_j) = \frac{Tweet_i\ .\ Tweet_j}{\|Tweet_i\| * \|Tweet_j\|} \tag{4}$$

Experimentally we found the maximum cosine similarity threshold which is used to determine whether two tweets should be linked or not. Tweet pairs with cosine similarities higher than or equal to 0.063 are linked and we obtained a connected graph.

Finally we assigned initial scores to the train nodes of the graph according to the results of the first phase. Similar to studies [22] and [18], nodes have two kinds of scores namely hub and authority. Authority score of a node indicates the direct meaningfulness value to the examined dimension while hub score shows the degree of being connection point among meaningful nodes. During iterations hub scores are used to calculate authority scores. Similarly authority scores are used to update hub scores as well.

4.2.2 Random Walk Iterations On The Graph

10 fold cross validation is applied in graph based improvement phase as well. A tenth of tweets are separated as test set while the rest of the tweets and all of the user nodes are assigned initial scores. Positively classified tweets in the first phase are assigned 1000 and negatively classified tweets are assigned -1000 initial hub and authority scores. On the other hand, users with credible tweets assigned 1000 hub and authority scores and the rest of the users are assigned -1000 hub and authority scores.

After constructing the graph and assigning initial hub and authority scores to the nodes, we run a predefined number of iterations on the graph for hub/authority transfers among nodes. At the end of those iterations, a tweet is classified as positive if its final authority score is greater than zero, and classified as negative otherwise.

$$Node\ N_j's\ hub\ score = \sum_{i}^{linked\ nodes\ with\ N_j} weight * N_i\ authority\ score \tag{5}$$

Hub score of a node is updated by adding a predefined ratio of authority scores of the neighbour nodes to its hub score according to Equation 5.

$$Node\ N_j's\ authority\ score = \sum_{i}^{linked\ nodes\ with\ N_j} weight * N_i\ hub\ score \tag{6}$$

Authority score of a node is updated by adding a predefined ratio of hub scores of the neighbour nodes to its authority score according to Equation 6.

Nodes' hub and authority scores increases or decreases according to the link structure of the graph during the random walk iterations.

4.3 Dictionary Based Analyses

For analysing the being free from offensive words, we tried some slightly different approaches as well. We checked the existence of offensive words from a slang-word dictionary. Other than the explained hybrid method, we made 3 more experiments for the first dimension of credibility:

1. Only considering word existence in slang-word dictionary of tweet text
2. Considering both word existence in slang-word dictionary of tweet text and first phase classification result
3. Selecting the tweets with negative sentiment score less than -2 which also contain slang word

Those methods are used to make the initial classification of the tweets. After the initial hub/authority score assignment according to each one of those methods, random walk iterations are applied and hub/authority scores are transferred in the graph. Performance results of those 4 methods are compared in the section 5.

5 EXPERIMENTS AND RESULTS

Three questions are asked to volunteers and majority of votes are used to determine three dimensional ground truth statuses of the tweet. Each question defines a credibility dimension of this study and each dimension is experimented separately. In this section we compared the classification results of the proposed hybrid method with human vote based ground truth data.

We made a large number of experiments. In this section we show the best results obtained. Best supervised learning phase results are obtained with j48 decision tree classification algorithm. We used WEKA API for machine learning algorithms. Second phase experiments are conducted with different hub/authority weights and best yes class recall(YCR), no class recall(NCR), yes class precision(YCP), no class precision(NCP) and F1 score(F1) of the experiments are shown in the figures.

5.1 First Dimension - Being Free From Offensive Words

The first dimension is about filtering offensive tweets. To check this, in the user study, volunteers were asked the following question: "Does the tweet contain swearing, abusing or offensive words?"

Table 2. First Dimension Supervised Learning Phase Best Results

Yes Class Recall:	0.351
No Class Recall:	0.978
Yes Class Precision:	0.457
No Class Precision:	0.966
F1 Score:	0.397
Accuracy:	0.947
Specificity:	0.978
Sensitivity:	0.353

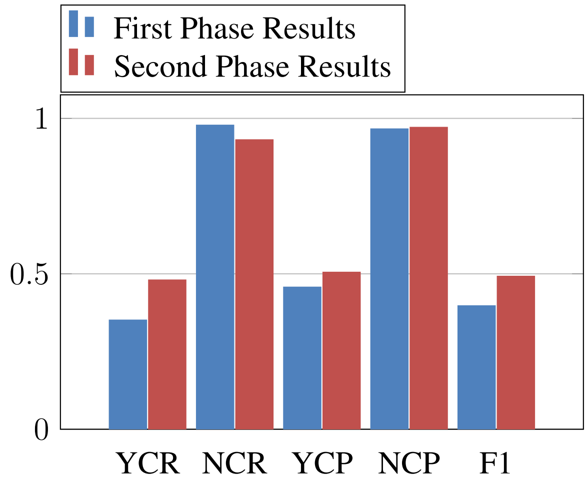

Figure 2. Random walk iteration improvement results of the first dimension.

In Figure 2, during second phase, YCR results increased 36% and YCP results increased 10%. NCR results decreased 4% and NCP slightly changed. Final F1 score increased 24% at the end of the graph based improvement phase.

We also made slang-word dictionary based experiments for this dimension. Those experiments and their performance results are given below:

- E1: Proposed hybrid method without dictionary based improvements
- E2: Only considering word existence in slang-word dictionary of tweet text
- E3: Considering both word existence in slang-word dictionary of tweet text and first phase classification result
- E4: Selecting the tweets with negative sentiment score less than -2 which also contain slang word

Table 3. First Dimension Slang-Word Dictionary Based Methods' Performances

Experiment:	E1	E2	E3	E4
YCR:	0.351	0.373	0.787	0.413
NCR:	0.978	0.942	0.934	0.966
YCP:	0.457	0.252	0.384	0.392
NCP:	0.966	0.966	0.988	0.969
F1 Score:	0.397	0.301	0.516	0.403

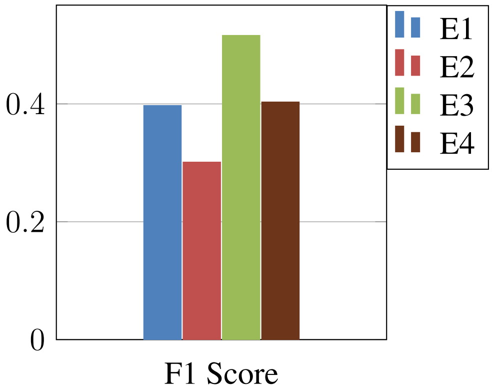

Figure 3. First Dimension Slang-Word Dictionary Based Methods' F1 Score Comparisons

As it can be seen from Figure 3, best F1 score is obtained in experiment 3. Selecting tweets with slang-words and considering supervised classification results together increased the F1 score performance 30% with respect to original method for first dimension analysis.

5.2 Second Dimension - Being Free From Spamming

The second dimension is about filtering spam tweets. To check this, in the user study, volunteers were asked the following question: "Is the tweet written for distracting, unrelated, advertising or out of program scope purposes?"

Table 4. Second Dimension Supervised Learning Phase Best Results

Yes Class Recall:	0.569
No Class Recall:	0.946
Yes Class Precision:	0.464
No Class Precision:	0.936
F1 Score:	0.511
Accuracy:	0.897
Specificity:	0.946
Sensitivity:	0.569

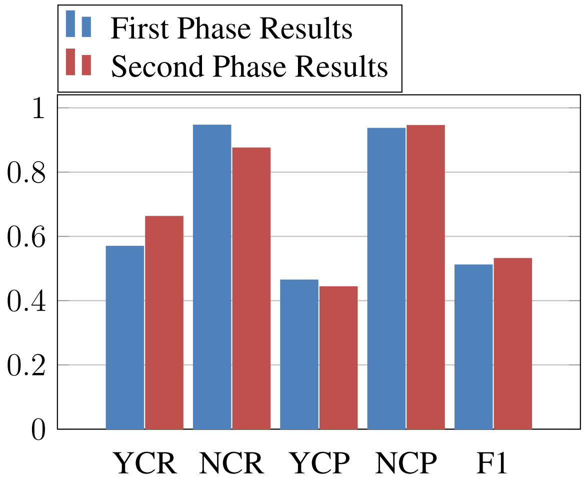

Figure 4. Random walk iteration improvement results of the second dimension.

In Figure 4 we observed that YCR increased 16% but YCP decreased 5%. On the other hand NCR decreased 8% whereas NCP incrased 1%. The graph based phase improved final F1 score 4%.

5.3 Third Dimension - Being Newsworthy

The third dimension is about the newsworthiness. To check this, in the user study, volunteers were asked the following question: "Is the content interesting, important or newsworthy?"

Table 5. Third Dimension Supervised Learning Phase Best Results

Yes Class Recall:	0.843
No Class Recall:	0.623
Yes Class Precision:	0.832
No Class Precision:	0.642
F1 Score:	0.837
Accuracy:	0.775
Specificity:	0.623
Sensitivity:	0.844

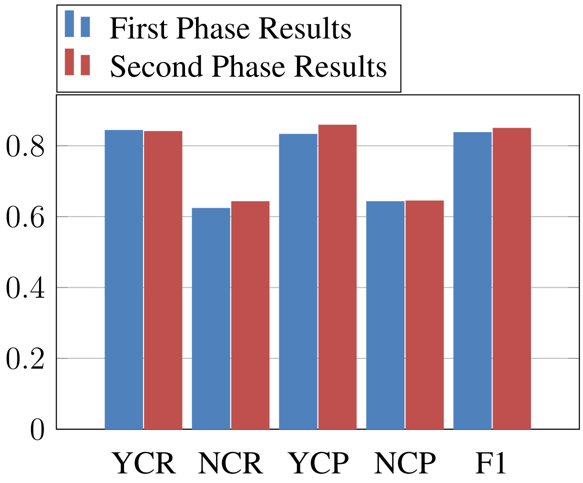

Figure 5. Random walk iteration improvement results of the third dimension.

In Figure 5 YCR and NCP changed insignificantly less than 1% but NCR and YCP increased 3%. Final F1 score is improved only 1%.

6 CONCLUSIONS

In this study we proposed a new method to analyse tweet credibility by hybridizing content based techniques with collaborative filtering based techniques. The proposed method consists of two phases. In the first phase we applied classification algorithms on the tweet set. 22 features of tweet text are obtained and used in the supervised learning phase. In the second phase we constructed a connected graph from users and tweets in which the edges are created according to author-text relation between users and tweets, friendship relation between users and normalized contextual similarity between tweets. We investigated whether credible tweets are linked with each other or not. We aimed to separate positive and negative classes by applying hub/authority score transfers in the graph.

We brought a new credibility definition based on three dimensions: being free from offensive words, not being spam and being newsworthy. Those three dimensions are examined separately in supervised learning and graph based improvement phases.

This study focused on the tweets written in Turkish language. We created our data set from tweets written for current Turkish TV programs about social and political discussions. Even though we developed a method based on Turkish language, the proposed method can be generalized for other languages by changing language parser and word separator components.

REFERENCES

[1] B. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87. ACM, 1999.

[2] B. Kang, J. O'Donovan, and T. Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 179–188. ACM, 2012.

[3] B. Kang, T. Höllerer, and J. O'Donovan. Believe it or not? analyzing information credibility in microblogs. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 611–616. ACM, 2015.

[4] B. J. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723. ACM, 2003.

[5] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[6] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *arXiv preprint arXiv:1407.5225*, 2014.

[8] M. Forelle, P. Howard, A. Monroy-Hernández, and S. Savage. Political bots and the manipulation of public opinion in venezuela. *arXiv preprint arXiv:1507.07109*, 2015.

[9] A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE, 2010.

[10] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[11] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Recent Advances in Intrusion Detection*, pages 318–337. Springer, 2011.

[12] J. Martinez-Romo and L. Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000, 2013.

[13] E. M. Clark, J. R. Williams, R. A. Galbraith, C. M. Danforth, P. S. Dodds, and C. A. Jones. Sifting robotic from organic text: A natural language approach for detecting automation on twitter. *arXiv preprint arXiv:1505.04342*, 2015.

[14] M. Mccord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing*, pages 175–186. Springer, 2011.

[15] I.-A. Bara, C. J. Fung, and T. Dinh. Enhancing twitter spam accounts discovery using cross-account pattern mining. In *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pages 491–496. IEEE, 2015.

[16] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[19] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *Web Information Systems Engineering–WISE 2010*, pages 240–253. Springer, 2010.

[20] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 564–575. VLDB Endowment, 2004.

[21] A. Gün and P. Karagöz. A hybrid approach for credibility detection in twitter. In *Hybrid Artificial Intelligence Systems*, pages 515–526. Springer, 2014.

[22] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *COMPUTER NETWORKS AND ISDN SYSTEMS*, pages 3825–3833. Elsevier Science Publishers B. V., 1998.