Automatic Text Summarization Methods Used on Twitter

Nazan Kemaloğlu[1], Ecir Uğur Küçüksille[1]

[1]*Department of Computer Engineering, Süleyman Demirel University,*
West Campus, 32000, Isparta, Turkey

Abstract—**Automatic Text Summarization; is one of the areas of Natural Language Processing which has become very popular especially in recent years. Generally automatic text summarization; is the process of getting a summary of a document given as input to the computer as output. The documents used for summarization are usually selected from news texts, corner texts or research texts. In addition to this, efforts are being made to achieve the same successes on documents that are defined as microblogs and that appear to be relatively short and meaningless. In this study, automatic text summarization methods used on the data obtained from Twitter, which is one of the most widely used microblog sites today, are examined. Summarization performances were evaluated in the light of the obtained findings, the methods used were examined and the difficulties encountered and their solutions were presented.**

Keywords—**Natural Language Processing, Automatic Summarization, Social Media Analysis, Twitter**

I. INTRODUCTION

Today's developing technology brings new problems along. The needs to solve these problems lead to the emergence of new research areas and computer science. Natural Language Processing in this context; It was born from the need to tell to the computers, the language used by people, and the language to be solved in the most accurate way, and it is a study field and computer science that has become very popular especially in recent years [1]. Thus, natural language processing has become more intense in recent years. Today, many problems such as text spoken, spelling, spelling mistakes, extracting information, summarizing, understanding, translation and question answering are tried to be solved by natural language processing [1] [2]. Due to Twitter's 140-character limit, users have developed as much information fitting strategies as possible on tweets they have written [3] [4]. Since Twitters inception in 2006, it has grown at an unprecedented rate. While the majority of tweets that are sent by users, are pointless babble or conversational, approximately %3.6 of these posts are topics of mainstream news, and another %8.7 are topics interesting enough for users to forward to their own followers via re-tweeting [5]. With approximately a million news posts being sent a day as well as many other sources of information, Twitter has become an important source of gathering real-time information on almost any topic imaginable. For most topics, users are forced to read through related posts (by clicking on the topic) in order to

try and understand why a topic is trending. This process is tedious and error prone as returned posts are prioritized only by recency. Therefore, for a given topic, users are likely to encounter spam, posts in other languages, rants, and other sources of misinformation. This problem can be solved by a summarization system on twitter [6].

II. AUTOMATIC TEXT SUMMARIZATION

Auto Text Summarization; is the process of summarizing a text through a computer program. This process can be briefly described as a procedure in which a summary document is obtained as an output element, from a text given as an input element to a computer program. This summary, should contain all the important points of the text and should be meaningful [7]. The summarization can be done in two ways; First method is to select the important sentences in the text by using the scoring methods, statistical methods and intuitive approaches, and to generate the summary with these existing sentences. The second method is the abbreviation of the sentences in the text and the interpretation of the entire text. There is a need for a rich charts of symbolic words to use this method [8] [9].

III. TWITTER

Twitter has been around since 2006; the activities that people make instantly; that allows sharing of economic, social, environmental or political situations, considerations, information on various subjects and news with the help of a web browser, SMS, e-mail or 3rd party applications, is defined as the world's most popular microblogging network, which houses a different user [10] [11].

IV. RELATED WORKS

Shankar, Karunamoorthy and Bhayani in 2015; have taken tweets from topics that are trend topic. They cleaned the collected tweets before the summarization process. In this process;
- Emotional expression,
- Extra marks such as # , &, @,
- Repeated tweets were removed with RegExr.

The remaining tweets was named "clean data" and separated by words with the Natural Language Toolkit (NLTK). The words separated by NLTK are labelled according to their type with NLTK POS and the stop words was removed. The data set obtained after this step was called "pre-processed data". In the phase of the sentence selection, a system based on WordNet and Lesk Algorithm is designed. Here, the meaning of each word is obtained from WordNet, and the

LESK algorithm chooses the most appropriate one from these meanings. The sentence weight is increased for each overlap obtained and this process was performed for all tweets. Finally, the four sentences with the highest weight were selected to form the summary sentences under that hashtag. In order to compare the results, certain people selected 4 tweets for each hashtag. Each choice was compared with automatic summarization, by Semantic Similarity and Cosine Similarity methods. The rate achieved has remained at %40 and it has been suggested that the inputs that given to the system, should be of better quality in order to increase this success. It has been argued that success can be very high especially in news-based tweets [12].

Sharifi, Hutton and Kalita in 2010; For the tweet collection process, firstly a tag was received from the user and tweets with a total of 50 titles under this tag were collected. In these tweets, non-English, retweets and spam were removed. For spam detection, a Bayes Classifier previously trained with spam data was used. Besides this, more than one tweet that sent by the same user was removed

from the collected tweets. Phrase Reinforcement (PR) expression enhancement algorithm was used in the summarization for edited tweets. This algorithm works according to the graph theory established around the determined hashtag. The keyword is placed as a center node and the words of sentence as leaf nodes are connected in the form of graphs. Each node was rated according to the frequency of expression and a summary was created with the highest score was selected. Here, nodes with center node and stop words are rated as "0".

According to Fig. 1; selected sentence as a result of scoring was: "A legend gone: Ted Kennedy died.". In order to measure the success of the extracted summaries, 2 persons were asked to summarize the same dataset. It was then desired to give a score of 1 to 5 (no-high) to the similarity of an independent observer's automatic and human-made summaries. As a result, it was seen that if the found road expresses the full dataset and the dataset was compatible with the selected title, the summarized result was meaningful. On the other hand, when it is seen that it is usual to select the long path according to the structure of the algorithm, the penalty point was added as it moves away from the center node [13].
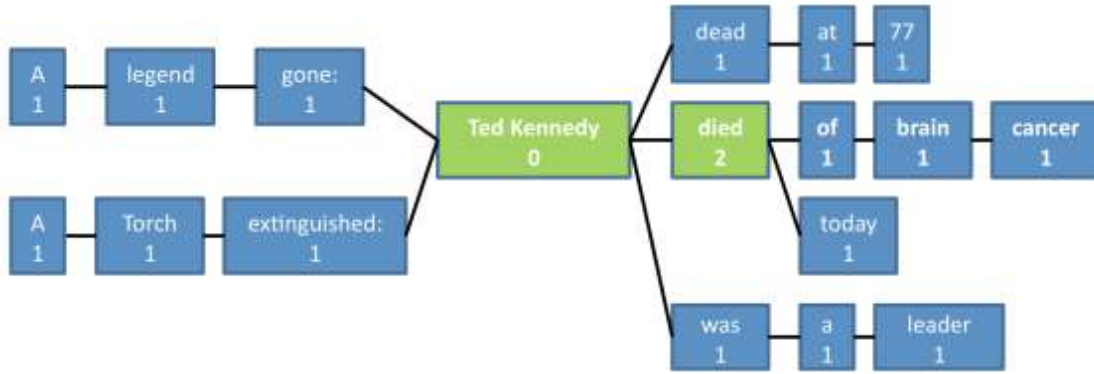

Fig. 1. Scoring by PR Algorithm

Sharifi, Hutton and Kalita in 2010; They used Hybrid TF-IDF algorithm besides PR algorithm in addition to previous studies. The TF-IDF algorithm that termed as term frequency, is calculated according to Equation (1).

$$TF - IDF = tf_{ij} * log_2\left(\frac{N}{df_j}\right) \tag{1}$$

Here ; is the frequency of the term T_j document of D_i N is total number of documents, df_j is the total number of documents containing the term T_j. Hybrid approach combines the sentences in each document and treats them as a single sentence. The disadvantage of this approach is that it captures high scoring in long sentences. Sentence scores were normalized to prevent this situation. Besides, the stop words were also determined and the weight was calculated as 0 in each case. As a result, the Hybrid TF-IDF algorithm can be used as much as the PR algorithm. As a result of the evaluation made, it has been observed that the TF-IDF algorithm reduces the success of the method by the long sentence sensitivity feature [14].

Kebabcı in 2015; collected tweets about earthquake. In 2014 and 2015, tweets about earthquakes in California, Nagano and Nepal were reviewed. Simplification and correction processes were done on the collected tweets:

- Repeated tweets were removed,
- The non-English tweets were removed.
- The words "with", "up", "to" which were not meaningful alone in English, were removed.
- The usernames that starting with expression were removed.

The dataset obtained is labeled as low and high priority. This labeling was made by hand; tweets with injured and damaged status information was labeled as high priority. Later, the high and low priorities of these tweets were classified by Naive Bayes, Support Vector Machines and Random Forest Algorithms. Summarization was performed on high priority tweets. The summarization methods that used were SumBasic, TF-ISF and Hybrid TF-IDF. In the SumBasic method, firstly the frequencies of the words used in the set are calculated. The summary score of that tweet is then calculated by adding the frequencies of the words in each tweet. The k tweets with the highest calculated value are selected as summary tweets [15]. In the TF-ISF (Term Frequency - Inverse Sentence Frequency) method, the frequencies of the words used in the set are calculated as in SumBasic. The summary score is then calculated using the general frequency values of the words in each tweet and the frequencies in the tweets of the words. The k tweets with the highest calculated value are selected as summary tweets

[16]. As a result of the work done, it was seen valuable tweets that came to the forefront in terms of information including dead, injured, and damage information. The tweet points in SumBasic are calculated by summing the use frequencies of the words. So the long tweets are more prominent. In the TF-ISF method, because of the frequency of use of the words in the tweet, the tweets containing the repeating words are foreground. Unlike Hybrid TF-IDF's TF-ISF, adding more general tweets to the account and Unlike Hybrid TF-IDF's SumBasic, including normalization process; has yielded more successful results. The summary of the work in Fig. 2 was given [17].
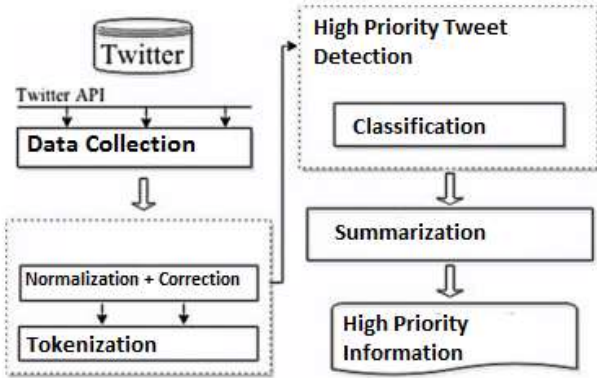

Fig. 2 The schema of summary process

Rosa,Shah, Lin, Gershman and Frederking in 2011; have examined the tweets that obtained in four different ways. These are named V0, V1, V2 and V3.

- V0: All words were written in small letters with spaces.
- V1: The spaces between the words were left out and lowercase letters are used, and terms with less frequency than 5 in all the dictation are removed.
- V2: In addition to V1 features, all characters except the special characters such as #, @, & were removed.
- V3: In addition to V2 features, URL and User names are taken as two separate classes.

These 4 sets were clustered to 30 and 6 respectively with educated and untrained clustering methods. Supervised methods; SVM, KNN, LDA and k-Means from unsupervised clustering algorithms. For each cluster created after clustering, it was generated a summary that will represent that cluster. The TF-IDF was used for the summarization algorithm. As a result, while comparing the

automatic summary with the man-made summary; precision, F-score and recall were used.

$$Precision(K) = |S \cap T|/S$$

(2)

$$Recall(A) = |S \cap T|/T$$

(3)

$$F - Score = 2KA/K + A$$

(4)

Where S is automatic summary's sentences, T is the sentence of the human hand summary. Moreover, it has been found that the TF-IDF was given more effective results in the unsupervised clusters [18].

Inouye and Kalita in 2011; were collecting 150 tweets that are taken under 10 topics that are trending consecutively for 5 days. So, they have obtained 1,500 tweets under 50 topics in total. These tweets were collected in HTML format have the following pre-processes applied:

- Each HTML encoded character was converted into ASCII code.
- Each Unicode character was replaced with ASCII code table counterpart, or it was deleted if not.
- URLs, HTML tags, headers (NEWS, etc.) was deleted.

The pre-processed data set has been tried to be summarized by 8 different summarization methods. Hybrid TF-IDF was the dwell on method of summarizing in the study. The remaining 7 methods were used to observe the accuracy of the Hybrid TF-IDF method. The data to be compared with summaries for accuracy checking for these 8 summarization algorithms used were obtained from independent users with Amazon Mechanical Turk. With Amazon Mechanical Turk, users were asked to choose their preferences by giving tweet to the system and the most selected ones were determined and a human fifty summary was created. For the evaluation of the summarization, the ROUGE-N package was used. ROUGE is a software developed to find similarities between two summaries. Where N; is n-gram. n-gram, is used in natural language processing, probability, theory of communication and data compression. It is a statistical model based on the prediction of what the next word (or letter) is. All ROUGE metrics contain three attribute values. These properties are Precision (2), Recall (3) and F-Score (4). Results of precision, F-score and recall are given in Fig 3.
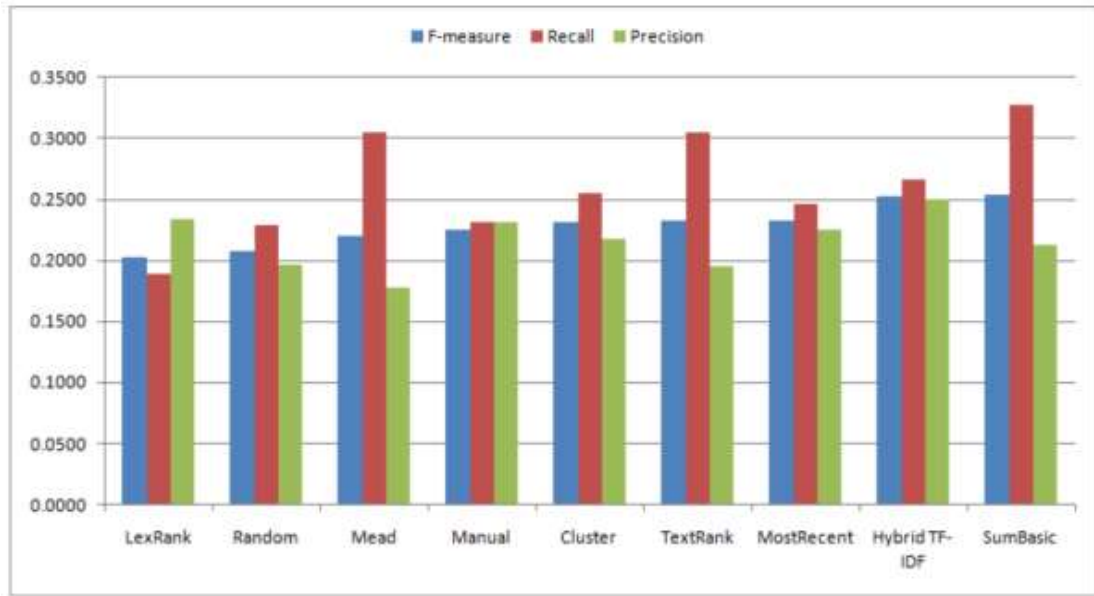
Fig. 3 Average F-measure, precision and recall ordered by F-measure

.Hybrid TF-IDF and SumBasic algorithms have been showed the best results from the summarization algorithms. It is emphasized that the most efficient summary method that can be used on Twitter is the term frequency based methods [19].

Kebabcı and Diri in 2017; were worked on Turkish tweets. The collected tweets were obtained from the official Twitter accounts of the 10 municipalities of the districts of Istanbul. Collected tweets were classified first, then extracted summary tweets that would represent each class. The summary of the work in Fig 4 is given.

Fig. 4 System Design

Collected tweets were cleaned first by pre-processing.

- User names, punctuation marks and words beginning with numbers were removed from tweets.
- The "#" character was removed.
- All letters were converted to lowercase.

Then the abbreviations used were checked and corrected. For example, the abbreviation "tsk" was corrected as "Tesekkur ederim", mean is "thank you" in English. It was also examined whether this abbreviation is an abbreviation used in Turkish language. When "tsk" was corrected, it was checked not to be confused with the TSK, (Turkish Armed Forces) in Turkish. In addition, the repetitive letters (eg: coook) were cleaned. For these operations lexical dictionary consisting of Turkish terms was used. If the abbreviation used was not found in the dictionary, it was removed. The tweets collected after the preliminary actions were labeled as "positive", "negative", "unknown", "question", "request" and "information / news" by two people. In the classification phase, SVM and Bayes classifiers were preferred. Tweets were first tried to be classified with SVM and the most probable 2 classes were determined and reclassified with Bayes classifier that trained by possible two classes. An average of %68 success was achieved in the classification process. For each class after classification; the summary to represent that class were tried to be found. For this, the Hybrid TF-IDF algorithm that described previously was preferred. With this method, 3 tweets with the highest score were selected as summary. As a results, only one summary tweet which has highest score is given in Table I..
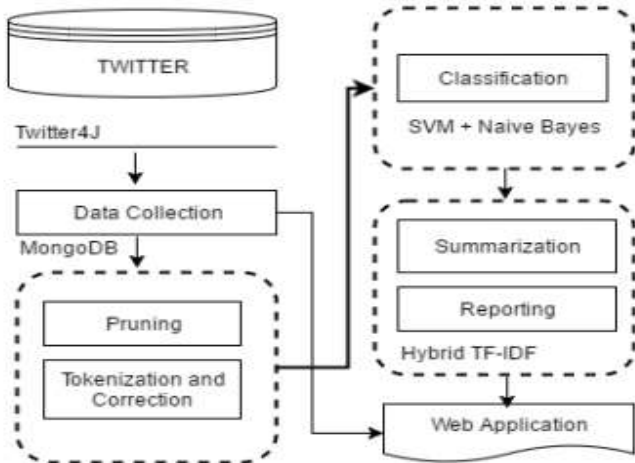
TABLE I. SUMMARIZATION RESULTS

Class	Summary Tweet
Positive	Tam anlamiyla sokak iftari... Tesekkurler @uskudarbld su güzel organizasyon icin tebrikler @hilmiturkmen34! (it was exactly street iftar... Thanks @uskudarbld for this good organization, congratulations @hilmiturkmen34!)
Negative	Sizin ecdat sevginiz işte bu kadar! RANT, RANT, RANT @uskudarbld @hilmiturkmen34 (That's all your ancestry love! RANT,RANT,RANT @uskudarbld @hilmiturkmen34)
Question	@uskudarbld Kız kulesine giden yolda sahildeki korkuluklar ne zaman yapilacak? Bu kadar ihmarkarlik olur mu????? (@uskudarbld When will the fence on the beach which is on the road to the Maiden's Tower be put? Would it be so negligent?)
Request	@hilmiturkmen34 @uskudarbld maganda şöförler çalıştırmak yakışmamaktadır. Gereğini yapmanızı rica ederim. İyi çalışmalar. (@hilmiturkmen34 @uskudarbld employment of roughneck drivers isn't befit you. I wou
Information /News	Acibadem TiBAŞ Parkı önünün @uskudarbld since 30 yıllığına Eğitim ve Sosyal amaçlı kullanımı için oylama bugün! (The voting for Acibadem TIBAS Park's use for Education and Social use for 30 years by @uskudarbld today!)

According to this study; It has been argued that summarization results are relative and summaries that are manually reviewed are acceptable for this reason [20].

Nichols, Mahmud and Drew in 2012, were taken account of the time factor for summarization. Sports games were preferred because the maximum of tweets can be collected instantly while generating the dataset. For this reason, the complete tweets collected for the 3 games collected were used as dataset. Tweets were collected during the competitions (Fig 6). According to this paper, tweets are more important at the time of increasing the status update volume. Also time elements have been used to clean spam or irrelevant topic tweets.

This type of tweet was set to be noisy and cleaned with noise reduction algorithms. Firstly; for noise reduction process with the help of the Apache Nutch project, the non-English tweets were removed. Second, using a glossary of common terms in spam data, it was the detection and removal of spam tweets. As a final step, normalization operations were performed on the words. For example, the "goooal" word was corrected as a "goal". In the summarization process, the sentence selection was made by Sharifi and others [13] [14] using the graph-based scoring method that previously mentioned. According to Fig 7; the phrase graph that would be generated from these status updates: *Landon Donovan scores! Donovan scores a brilliant strike, Landon Donovan is brilliant.*
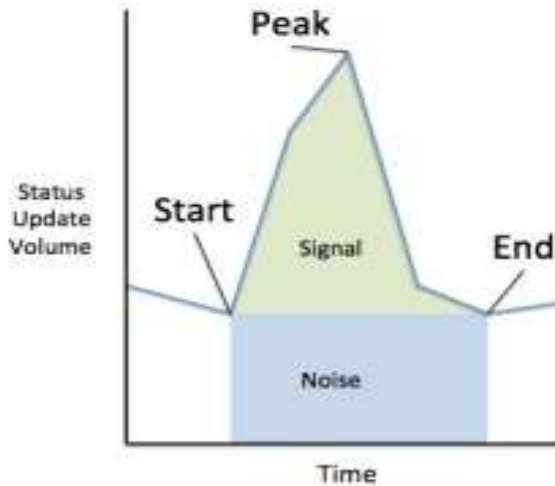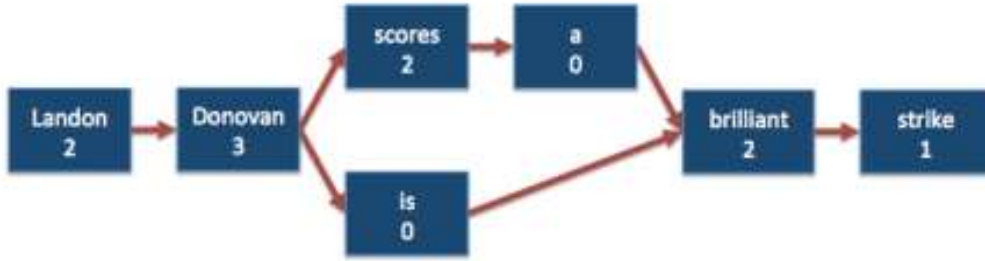

Fig. 6 The anatomy of times

Fig. 7. An example phrase graph. Words and their respective weights are shown in each box.

For the evaluation of the summarization, the ROUGE-N package was used. Two summaries were used for the evaluation. These were a summary of tweets made by a team of 6 people and the second was the summaries of the FIFA.com. Also TF-IDF, which is the general method of summarization, was used to evaluate the accuracy of the results. Results of precision, F-score and recall are given in Fig 8 and Fig 9.
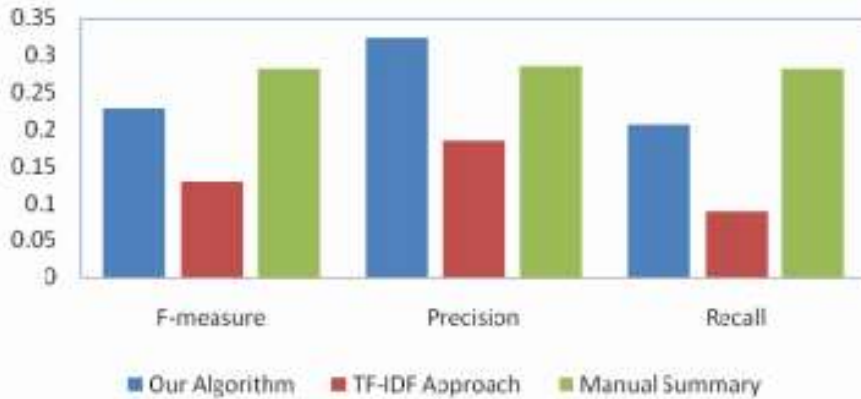

Fig. 8. ROUGE-1 comparison of our algorithm, TF-IDF algorithm and manual summaries for human evaluator
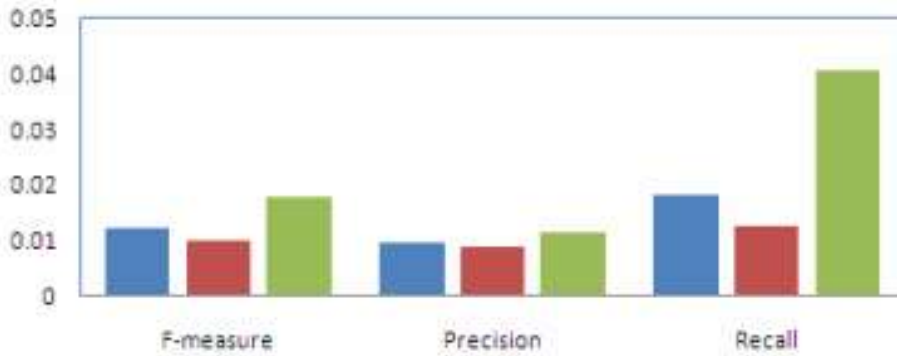

Fig. 9. ROUGE-2 comparison of our algorithm, TF-IDF algorithm and manual summaries for FIFA.com Summaries.

According to Fig 8 and Fig 9; ROUGE 1 shows that the algorithm developed was sufficient, ROUGE 2 comparison was low. The reason for this the summaries from FIFA.com were relatively longer than our summaries. At the same time, it was seen as an important reason why authors and twitter users not use the same words. However, the tweets received in the study were neutral tweets. How the performance of the system will be affected if the collected tweets are in favour or against any team is also determined as a future work to be done [21].

V. CONCLUSION

When reviewing the Tweet summarization literature, it is seen that the studies conducted are carried out on tweets collected in English language mostly. Besides this, all studies; has shown that a pre-process must first be performed on the tweets. Since the tweets to be used are not on a certain level, it is necessary to bring the same form in order to use. In addition to this, it has been determined that the titles used in the summary are of great importance, and that the titles used in the news or the research field mostly contain favorable tweets to produce a more meaningful summary. In the summarization phase, graphs and statistical methods have been preferred as well as clustering methods such as k-means and LDA. When the literature was examined, it has been determined that statistical methods based on the term frequency give the most efficient results for summarizing. In addition to this, summarizing the

summaries of the tweets that were collected under a certain heading into subheadings seems to contribute to the increase in success. In addition to this, it has been seen that all the summarization methods on tweets were chosen among presumptive based methods. Summarization has been done in the form of sentence selection from the tweets obtained. This is due to the fact that summarizing the tweets made up of interpretation-based clauses with interpretation-based methods will not yield a meaningful result. Most preferred method was TF-IDF for summarization. In order to determine the correctness of the summarization process, a complete standard has not yet been established. Most of the verification procedures are based on summary by independent reviewers of available tweets. The similarity ratios of the summaries issued by the system and the abstracts of the persons indicate to us the correctness of the system. The most commonly used methods in the literature, was ROUGE. As a result, it was seen that the tweet summarizing process did not give accurate and descriptive results like document summarization. It is anticipated that a user-assisted summarization process will increase success for semantic accuracy. For summarization, it will be considered that the selection of sentences containing the keywords that the user will give to the system will be more accurate.

REFERENCES

[1] Coban, O. "Turkish Twitter Sentiment Analysis Using Text Classification Techniques". PhD thesis. Ataturk Universty, 2016.

[2] Adali, E., "Natural Language Processing". In: Journal of Turkey Informatics Foundation of Computer Science and Engineering 5.6 (2012).

[3] Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In Proceedings of the 2012 international workshop on Smart health and wellbeing (pp. 25-32). ACM.

[4] Meral, M., & Diri, B. (2014). Sentiment analysis on Twitter. In Signal Processing and Communications Applications Conference (SIU), 2014 22nd (pp. 690-693). IEEE.

[5] Schonfeld, E., Pingdom Says People Are Tweeting 27 Million Times A Day. 2009. url: https://techcrunch.com/2009/11/12/twitter- 27-million-tweets-day-pingdo"(visited on 04/20/2018).

[6] Sharifi, B. "Automatic Microblog Classification and Summarization". PhD thesis. University of Colorado, 2010.

[7] Guran. A.,"Automatic Text Summarization System". PhD thesis. Yildiz Technical Universty, 2013.

[8] Uzundere, E., Dedja, E., Diri, B., Amasyah, M. F. (2008), "Türkçe Haber Metinleri için Otomatik Özetleme", In Proceedings of ASYU 2008, Isparta, Türkiye.

[9] Gundogdu, O., E., Duru, N. Methods Used in Turkish Text Summarization. 2016. url: http://ab.org.tr/ab16/bildiri/36.pdf (visited on 04/20/2018).

[10] Han., L., C. "Detecting News Topics from Microblogs Using Sequential Pattern Mining". PhD thesis. Queensland University of Technology, 2014.

[11] Gunduz M., S. "Classification of Content of News Posts on Social Networks". PhD thesis. Gazi Universty, 2014.

[12] Shankar, Y., Karunamoorthy, S., Bhayani, S.. An Approach to Automatic Trending Tweet Summarization. 2016. url: https://github.com/yuva29/twitter-trends-summarizer (visited on 04/20/2018).

[13] Sharifi, B., Hutton, M. A., & Kalita, J. (2010, January). Automatic summarization of twitter topics. In National Workshop on Design and Analysis of Algorithm, Tezpur, India (p. 4).

[14] Sharifi, B., Hutton, M.-A., and Kalita, J. K. (2010). Experiments in microblog summarization. In Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10, pages 49–56, Washington, DC, USA. IEEE Computer Society

[15] Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 43(6), 1606-1618.

[16] Wadhvani, R., Pateriya, R., K., Roy, D. "A Topic-driven Summarization using K-means Clustering and Tf-Isf Sentence Ranking". In: International Journal of Computer Applications 79.8 (2013), pp. 39{45}.

[17] Kebabcı, K. "Detection and Summarization of the High Priorty Tweets After Natural Disaster". PhD thesis. Istanbul: Yildiz Technical Universty, 2015.

[18] Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R. (2011). Topical clustering of tweets. In Proceedings of SIGIR Workshop on Social Web Search and Mining Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

[19] Inouye, D., & Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 298-306). IEEE.

[20] Kebabcı, K., Diri, B. "User Opinion Analysis on Turkish Tweets". In: Mugla Journal of Science and Technology 3.1 (2017).

[21] Nichols, J., Mahmud, J., Drews, C. "Summarizing sporting events using twitter". In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM. 2012, pp. 189-198.