The Effects of Emoji in Sentiment Analysis

Mohammed O. Shiha[1], Serkan Ayvaz[2]*

[1] Department of Computer Engineering, Bahcesehir University, Besiktas, Istanbul, Turkey.
[2] Department of Software Engineering, Bahcesehir University, Besiktas, Istanbul, Turkey.

* Corresponding author. Tel.: +90-212-381-0886; email: serkan.ayvaz@eng.bau.edu.tr

Manuscript submitted February 5, 2017; accepted March 1, 2017.
doi: 10.17706/ijcee.2017.9.1.360-369

Abstract: This study investigates the usage of Emoji characters on social networks and the effects of Emoji in text mining and sentiment analysis. As it provides live access to text based public opinions, we chose Twitter as our information source in our analysis. We collected text data for some global positive and negative events to analyze the impact of Emoji characters in sentiment analysis. In our analysis, we noticed that the utilization of Emoji characters in sentiment analysis results in higher sentiment scores. Furthermore, we observed that the usage of Emoji characters in sentiment analysis appeared to have higher impact on overall sentiments of the positive opinions in comparison to the negative opinions.

Key words: Emoji, opinion mining, sentiment analysis, twitter.

1. Introduction

Opinions and preferences expressed in social networks and microblogging services are essential in sentiment analysis and opinion mining. The social media sentiment analysis as an essential information source is currently being used for various purposes including spotting customers' dissatisfaction or problems with products; estimating stock market prices to even predicting the results of the elections before it takes place.

Sentiment in positive and negative terms can be considered as the most significant value of the opinions conveyed by the users on the microblogging services like Twitter. Sentiment Analysis (SA) uses Natural Language Processing (NLP) to extract opinions from users' text and classify them into classes in most cases negative, natural and positive. Such extracted opinions can be used by organizations or companies to help track products, improve services or predict upcoming events.

Twitter is one of the most substantial and popular social networks in the digital world. Approximately 695 million users tweet more than 58 million tweets daily. The number of active users exceeds 342 million as of September 2016 [1]. 43% of those users use their mobile phones to send their tweets. This makes Twitter even better as a source of sentiment data since the users can share their feelings or opinions about any event immediately as it occurs. Twitter users tweet about nearly all topics, which is also an advantage of Twitter. People share their feelings and opinions on negative events as well as positive experiences. Another important feature of Twitter is the use of hashtags. Users can use hashtags to categorize tweet topics. Hashtags help organize the topics and simplify searching for tweets about the topics. The hashtags can also be used to track tweets about a specific happening like "Oscars" or "World Cup".

Microblogging services and social media users use text to express their opinions or feelings. Twitter supports up to 140 characters long tweets to be posted at a time. The advantage of feature is that it constrains the users use a limited number of characters to express their opinions concisely. This allows the

Twitter data to be more normalized compared to other social networking sources.

The use of Emoji on the Internet rapidly increased in recent years. The ideograms and smileys enabled users express their emotions more easily the text in electronic messages and web pages.

Emoji such as "Face with Tears of Joy" have changed the way we communicate on the social networks and microblogging services. People often use them when it is more difficult to describe their expressions only with words. A single Emoji character may enhance the expressivity of a text message. A name of a city has no sentiment value when it is posted alone. However, if the user used an Emoji along with this name, the text may have a sentiment value. For example, a smiling face Emoji character "☺" can express someone's positive feeling towards the city. In contrast, using the angry face Emoji "😫" along with some brand name may reveal negative feelings towards the brand. An Emoji character can give a deeper meaning in a sympathy post. Another emoji may help show how happy the user is while trying a new drink.

A study published in the Social Neuroscience journal [2] showed that human brain reacts to Emoticons (a simpler form of Emoji) as real faces. Dr. Owen Churches, from the school of psychology at Flinders University, has found that Emoticons have become more important than we assumed. It appears that we now react to them in the same way as we would to a real human face.

In November 2015, Oxford Dictionaries Word of the Year was chosen to be an Emoji character which is known as "Face with Tears of Joy" or "😂" for the first time ever. This came after the wide usage for this Emoji character on the Internet, especially on social networks.

In this paper, we study the appearance of Emoji characters on social networks and how they affect the process of text mining and sentiment analysis. Events of positive and negative feelings and overall impressions have been studied in order to understand the nature and usage of such characters. We analyzed public opinions about global happenings as the main resource to investigate the effects and usage of Emoji characters on social network sentiments.

2. Background and Related Work

Sentiment Analysis models have been investigated by a large number of studies. The approaches and models vary across fields. However, only a small number of those studies investigated the usage of Emoji characters on social networks. The first Emoji lexicon was implemented by Novak *et al.* [3]. 83 human annotators were recruited to classify 1.6 million tweets collected in 13 different European languages into negative, neutral, or positive classes. They found that approximately 4% of collected tweets contained Emoji. Based on that they ranked the 751 most popular Emoji characters using the sentiment score of the plain text. To indicate the classes, they scored Emoji characters. In this study, we also developed Emoji sentiment scores in the lexicon that we established. F Barbieri *et al.* [4] retrieved 10 million tweets in English language to build skip-gram word embedding models by mapping words and Emoji in the same space.

Sentiment Analysis models have been developed for various purposes including predicting financial and economic variables, estimating customers' satisfaction, as well as some political aspects like predicting the results of the elections before voting. For instance, as a model of consumer brand sentiment, the study in [5] used a sample of 3516 tweets to evaluate consumers' sentiments towards leading brands. A predefined lexicon of 6800 seed adjectives with known orientation was used to conduct the analysis of the model. By using qualitative and quantitative methodologies to analyze the data, the model was able to indicate the consumer sentiments towards well-known brands.

MJ Rust, M Bates and X Zhuang in [6] used a multi-knowledge approach integrating WordNet, statistical analysis and movie knowledge to build an effective model in movie review mining and summarization. Wang *et al.* [7] proposed a real-time sentiment analysis system for 2012 US presidential elections expressed on Twitter. The system achieved 59% accuracy on four category classification by using Naïve Bayes model with unigram features on more than 36 million tweets collected.

Bollen *et al.* in 2011 [8] analyzed the text content of daily Twitter using mood tracking tools like Opinion Finder which measures positive and negative moods and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy) to predict the changes in DJIA closing values. The approach indicated that the accuracy of DJIA predictions can be improved by the inclusion of specific public mood dimensions. They achieved an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA stock market index.

Cheong, Marc, and Vincent CS Lee [9] proposed a sentiment model to collect public sentiment and response during terrorism scenarios. Accompanied with data mining, visualization, and filtering methods, the model yielded useful graphical visualizations of information, to reveal potential response to terrorist threats.

3. Emoji Lexicon

At the beginning, Emoji were used by a number of Japanese mobile operators. A set of 176 Emoji was first introduced as a message feature of a mobile operator at the time. In 2010, hundreds of Emoji characters were integrated with Unicode character set (Unicode 6.0) which grant Emoji the opportunity to be used worldwide. Reaching to Unicode 9.0, a total of 1,126 Emoji characters were supported.

Since Emoji characters are a subset of a large Unicode character set, they are represented and stored in the same way as all Unicode characters. Some websites came up with their own fonts to show Emoji characters in more appealing ways to attract more users. Furthermore, Unicode characters can have different values for their representations depending on the number system, programming language or even the notation itself.

Collecting data for Sentiment Analysis (SA) models differs based on the tools used to gather the data. Depending on the tool, researchers may receive a different form of the Emoji characters in the collected data. For example, Twitter API for R language provides these characters in different representations than Hadoop Flume service. In order to search for a specific Emoji character, the proper Unicode representation should be used in the software.

We prepared a lexicon containing 18 different representations for 843 different Emoji characters that supports Emoji notations including R, PHP, Java and other forms.

As the Sentiment Analysis of Emoji model [3] proposed an approach to give a score to Emoji characters through checking the sentiment score of the text itself and giving the Emoji character a score among values {-1, 0, 1}. The plain text was given to native speakers from the same language the text was written in. The speakers gave scores to the text, negative, natural, or positive. For each Emoji character the score of all text was aggregated and the average score was calculated. We derived the scores in our lexicon based on the scores developed by [3].

Additionally, Emojitracker.com [10] is a website designed to monitor the appearance of the Emoji characters on Twitter. The website tracks all public tweets and analyzes the presence of Emoji characters. The number of occurrences for each Emoji character is calculated, and sorted based on the number of occurrences. We developed a Java application to download the Emoji data. The application automatically parses the data and adds Emoji characters to our lexicon.

As a result, we built an Emoji Lexicon that supports up to 18 different notations and contains sentiment scores between 0 and 1 along with the rank of the Emoji on Emojitracker.com. This lexicon can help in parsing text and detecting Emoji characters for data collected from microblogging services or social networks. Table 1 shows an example of one Emoji character from the Lexicon "Face with Tears of Joy", the Emoji with the highest appearance on the Internet [11].

Table 1. Emoji Lexicon Sample Entry (Face with Tears of Joy)

Representation	Value
Emoji Name	FACE WITH TEARS OF JOY
UTF-8 Unicode Character(s)	😂
UTF-8 Character Count	1
Decimal HTML Entity	&#128514;
Hexadecimal HTML Entity	&#x1f602;
Hexadecimal Code Point(s)	1f602
Unicode Notation	U+1F602
Decimal Code Point(s)	128514
UTF-8 Hexadecimal (C Syntax)	0xF0 0x9F 0x98 0x82
UTF-8 Hexadecimal Bytes	F0 9F 98 82
UTF-8 Octal Bytes	360 237 230 202
UTF-16 Hexadecimal (C Syntax)	0xD83D 0xDE02
UTF-16 Hexadecimal	d83dde02
UTF-16 Decimal	55357 56834
UTF-32 Hexadecimal (C Syntax)	0x0001F602
UTF-32 Hexadecimal	01F602
UTF-32 Decimal	128514
Python	u"\U0001F602"
PHP	"\xf0\x9f\x98\x82"
C / C++ / Java	"\uD83D\uDE02"
R-Encoding	<ed><a0><bd><ed><b8><82>
Emojitracker.com rank	1
Sentiment Score	0.805100583

4. Model and Analysis

The use of Emoji on the microblogging services and social network increased significantly in the last few years. One of the most popular and influential social network platforms is Twitter. Similar to other social networks, the usage of Emoji on Twitter has increased. Unlike Instagram and many other social media platforms, the entries on Twitter are mainly text-based posts. Furthermore, the users typically use brief and concise sentences to express their feeling and opinions in their posts as Twitter limits the text size of each post to a maximum of 140 characters. Additionally, Twitter provides real-time access to public posts through its API. These features make Twitter a good platform for large scale real-time opinion mining. Due to these reasons, we chose Twitter as the primary data source amongst other social media for our analyses.

In order to evaluate the effects of using Emoji characters on the Sentiment Analysis models, the expressivity of the characters should be investigated. Unless punctuations and special characters are the main topic of the research, researchers usually remove all punctuations and special characters from the text data in text mining and sentiment analysis. Emoji characters are treated like any set of special characters and they are removed during the process of text mining. In this paper, our goal was to study the effects of Emoji characters on Sentiment Analysis models. We analyzed whether people use Emoji more frequently in positive or negative life events, and evaluated the impact of ignoring those characters while developing a sentiment analysis model.

4.1. Collecting Data

To evaluate the usage of Emoji characters in both positive and negative life events, we collected data from

Twitter on two important events occurred recently. The first one was "The New Year's Eve". New Year's Eve is a day of positive feelings for many people as they celebrate New Year's Eve to farewell the year as it ends and to welcome the New Year. People use social networks like Twitter to express their feelings, emotions and send their wishes to their families, friends, and loved ones. Thus, we selected New Year's Eve as event with positive feelings.

The second event was selected to be "Istanbul Attack" in which dozens of people were killed in a nightclub on December 31st in 2016 in Istanbul, Turkey as they were celebrating the New Year [12]. People also used social networks to express their sympathy and condolences with the Turkish people on one hand and talk about violence and terrorism on the other. This event was considered as a negative event.

4.2. Positive Event without Emoji

In order to study the usage of the Emoji characters in events with positive feelings, we chose the event "The New Year's Eve" as people tend to have positive feelings as they welcome the New Year. They celebrate this event in many ways including sending their wishes to each other using digital communications and social networks. Twitter as one of the most important social networks is considered as a place in which people can share their thoughts and wishes about the New Year. A total of 101,695 tweets have been collected using the hashtag "#HappyNewYear" starting from December 31st, 2016 for consequent five days. This hashtag is being used by Twitter users to mention "The New Year's Eve" event. Hundreds of thousands of Tweets were posted by Twitter users reflecting their feelings. 101,695 tweets were collected in English language from all over the world to make the approach as accurate as possible since our approach uses the natural English language words found in the tweets using English sentiment dictionary.

SentiWordNet [13] is a lexical resource for opinion mining. This dictionary contains two groups of English words: positive and negative.

SentiWordNet contains a total of 6,789 words, 4,783 positive words and 2,006 negative words. This dictionary is the main vocabulary lexicon used in our model. Comparing tweet's text with the dictionary, each positive word in the text is given a score of positive one (+1) and each negative word is given the score of minus one (-1). Words found in tweet's text and don't exist in the dictionary take the zero score by default which is objectivity score. For each tweet text, each word is replaced by the corresponding score from this dictionary. The sum of these scores indicates the score of the entire tweet. This way, we score the tweets and classify them into three classes: positive, neutral (objective) and negative classes. Positive words dictionary contains words represent the feelings and opinions like "good", "happy", "love" and "enjoy". Whereas the negative dictionary contains words that reflect the negative feelings and opinions such as "bad", "sad", "hate" and "terrorism".

We developed a Sentiment Analysis application in R language to score the tweets and classify them into positive, neutral or negative tweets. The application was prepared first to ignore all the punctuations, special characters, and Emoji characters too. In the first phase, the application processed the tweets collected for the positive event "The New Year Eve" with SentiWordNet dictionary. The application was designed to determine the score of each tweet by breaking it down to words and checking the sentiment score of each word in the sentiment dictionaries. The sentiment final score of each tweet was obtained by summing the sentiment score of each word in the tweet. In this phase, we calculated the sentiment score of the text itself without any influence from the Emoji characters.

A total of 101,695 Tweets were collected from Twitter about what people said about that occasion, and what the words that they used to express their feelings. English language was specified during the process of collecting the data. Among more than hundred thousand of tweets collected, around 19,714 tweets contained at least one Emoji character which gave us about 19.38% appearance in all the data related to the positive event. Fig. 1 shows the ratio between the number of tweets with and without Emoji characters for both events.

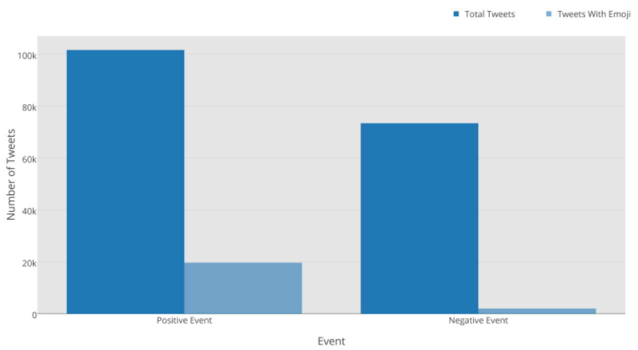

Fig. 1. Number of tweets collected for both events, light blue bars represent the total number of tweets collected. Dark blue bars indicates the number of Emoji characters found in each collection.

Scoring tweet process continued through multiple iterations. All special characters removed from the text since they don't give any sense of sentiment. Then, all punctuations have been removed. Like all Unicode characters, Emoji characters have been removed from the text.

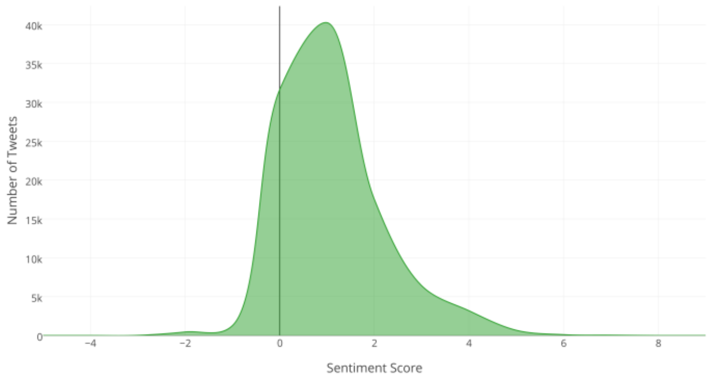

Fig. 2. Sentiment score of "The New Year's Eve" event without Emoji characters.

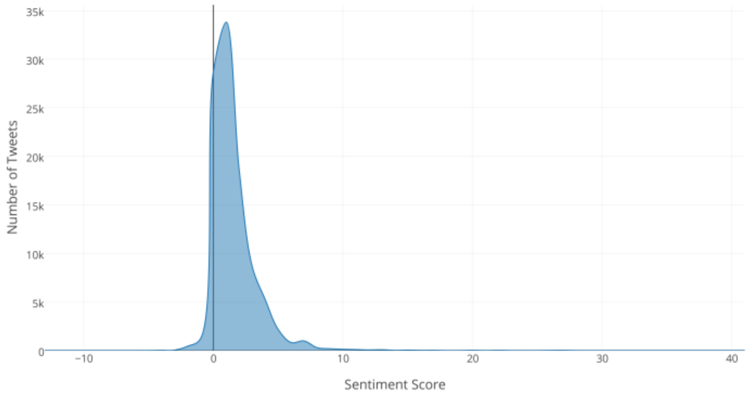

Fig. 3. Sentiment score of "The New Year's Eve" event with Emoji characters.

Fig. 2 shows the sentiment results obtained from scoring the text without the Emoji characters. 69,382 tweets (68.2%) classified as positive tweets, 30,555 tweets (30.0%) classified as neutral (objective) tweets, whereas only 1,758 tweets (1.8%) classified as negative tweets. We observed from the curve that the majority of tweets are on the right side of the zero line (objectivity line) which indicates that the event has a positive overall impression.

4.3. Positive Event with Emoji

More than nineteen thousand tweets appeared with Emoji characters in the positive event. In the first phase we ignored all of these characters by removing them from the text while scoring. In this phase Emoji characters are assigned a sentiment value. Sentiment values presented in the lexicon proposed by [3] determined the class of the Emoji characters. Emoji characters with positive sentiment score have been added to positive words dictionary. Emoji characters with negative sentiment score have been added to the negative dictionary. Emoji characters with sentiment score close to the objectivity haven't been added to any dictionary since they had the neutral score and didn't appear in any of the dictionaries.

We used two dictionaries to score the text in this phase. The first one is the positive SentiWordNet dictionary with the positive Emoji characters added to it. The second is the negative SentiWordNet dictionary with all Emoji characters with negative scores added to it. An Emoji character matching one of these dictionaries are classified accordingly and Emoji characters with no matches in any of the dictionaries took the neutral score (0).

We followed the same approach when analyzing sentiments without removing Emoji characters. Each positive Emoji character was given a score of (1). A negative Emoji character was given a score of (-1), and (0) was assigned to Emoji characters don't appear in any of the dictionaries.

Fig. 3 shows the results of scoring the data by taking Emoji characters in consideration when scoring. The positive results have been increased from 69,382 (68.2%) in the first phase (without Emoji) to 71,285 (70.0%) after scoring the Emoji characters. Neutral results decreased from 30,555 (30.0%) to 28,670 (28.1%), with no major change in negative results from 1,758 to 1,740 (1.7%).

An increase by approximately 2% in the positive scores of the tweets has been achieved. We also noted nearly 2% decrease in the neutral scores. Negative opinions kept the same value with 1.7% of all tweets. In other words, scoring Emoji characters changed the overall result by 2% in this case.

The figure demonstrates that the results stayed in the positive side. However, they took wider values on the scores axis. In the first phase (without Emoji) the maximum value was 8 and the minimum value was -4, comparing to this phase (with Emoji) the values took wider range starting from -10 to 40 which means that the text contains more sentiment scores than the first phase. One tweet can contain multiple positive or negative Emoji characters, or a combination of both.

4.4. Negative Event without Emoji

Twitter is a platform where people also show their sympathy and grief. On December 31st, 2016, coinciding with the New Year's Eve celebrations around the world, a terror attack happened in Istanbul, the biggest city of Turkey. Dozens of people were killed at the night. People around the world reacted to this event, and shared their opinions on social media. On Twitter, hashtags were used to group the tweets talking about this event, hashtags like "#IstanbulAttack" and "#PrayForTurkey" have been used by people on Twitter to express their feelings about it.

For this event, we used trending hashtags to collect data related with this happening. 73,486 tweets have been collected in English language. 2,064 of them contained Emoji characters as shown in Fig. 1. 2.8% of the tweets have Emoji characters which is a sharp decrease from more than 19% of Emoji usage in the positive event. SentiWordNet dictionary was used again to give score for each word in the tweets' text, and classify them into negative, neutral or positive classes.

The same approach was followed as in the positive event sentiment analysis. Tweet texts have been cleaned by removing punctuations, special characters and Emoji characters. Scores were assigned for each tweet in the data collected using SentiWordNet dictionary.

The results contained 46,967 tweets with negative scores, which means that the overall impression of this data was negative (around 64%), neutral results came with 20,056 tweets (27.2%) and 6,463 positive tweets (8.8%).

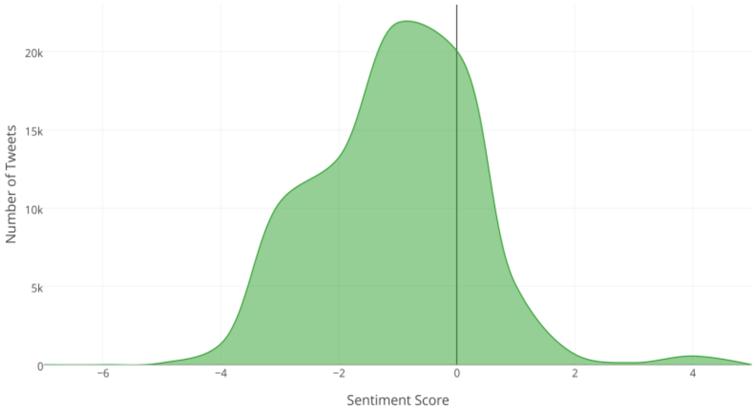

Fig. 4. Sentiment score of "Istanbul Attack" event without Emoji characters.

Fig. 4 shows the results for this phase. From the figure we found that the majority of tweets are on left side of the 0 line (objectivity line) which means that the results have more negative values than positive, especially that the positive results was less than 9% of the total opinions.

4.5. Negative Event with Emoji

A number of Emoji characters were added to represent the feelings of sadness, grief and anger. Emoji characters were also being used by people in the event with negative feelings.

In this part we took into consideration the existence of Emoji characters in the text during the sentiment scoring process.

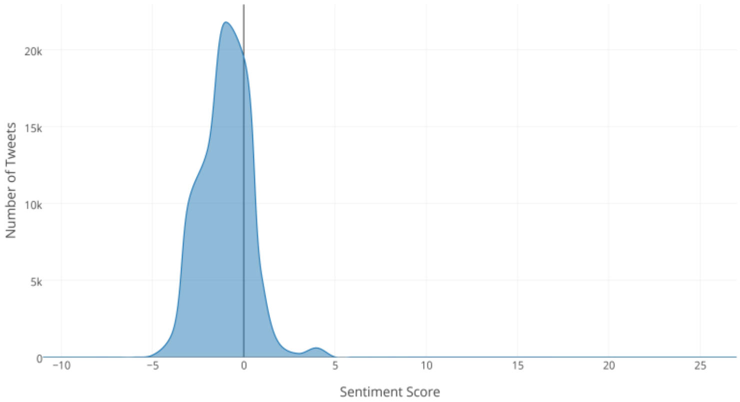

Fig. 5. Sentiment score of "Istanbul Attack" event with Emoji characters.

More than two thousand tweets in this data collection appeared with Emoji characters, which was relatively low number compared to the usage in the positive event. The appearance of these characters has been considered in this part. Fig. 5 shows the results of giving scores for Emoji characters in the negative event. We see that results were relatively close to the without Emoji characters. The negative results kept the same value as before, 47,064 tweets (around 64%), a decrease in the objective opinions from 27.2% to 26.6% (19,558 tweets), and a small increase in the positive opinions from 8.8% to 9.3% (6864 tweets). We also noted the wider range of scores. The analysis resulted in scores ranging between -10 and 25 whereas the analysis without Emoji was ranging between -6 to 4.

5. Conclusion

In this paper, we investigated the usage of Emoji characters on social networks and the impact of Emoji in text mining and sentiment analysis. We analyzed some global positive and negative events to find out whether there was a discrepancy of Emoji usage between positive and negative events. We observed that considering Emoji in sentiment analysis help improve overall sentiment scores. While Emoji characters are used for expressing both negative and positive opinions, the usage of Emoji characters in sentiment analysis appeared to improve the expressivity and overall sentiment scores of the positive opinions relatively more than the negative opinions in our analysis.

References

[1] Statistic Brain. (2016). STATS | Twitter company statistics. Statistic Brain. Retrieved Jan. 30, 2017, from http://www.statisticbrain.com/twitter-statistics/

[2] Churches, O., Nicholls, M., Thiessen, M., Kohler, M., & Keage, H. (Jan. 2014). Emoticons in mind: An event-related potential study. *Social Neuroscience, 9(2),* 196–202.

[3] Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (Dec. 2015). Sentiment of emojis. *PLOS One, 10(12).*

[4] Barbieri, F., Ronzano, F., & Saggion, H. (2016). What does this emoji mean? A vector space skip-gram model for twitter emojis. *Proceedings of Language Resources and Evaluation Conference, LREC,* Slovenia: Portoroz.

[5] Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40(10),* 4241-4251.

[6] Rust, M. J., Mark, B., & Xiaowei, Z. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods, 3(10),* 793-796.

[7] Wang, H., *et al.* (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics*.

[8] Bollen, J., Huina, M., & Xiaojun, Z. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1)*, 1-8.

[9] Cheong, M., & Vincent CS, L. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers, 13(1)*, 45-59.

[10] Emojitracker: Realtime emoji use on twitter. *Emojitracker.* Retrieved Jan. 30, 2017, from http://emojitracker.com

[11] Liberatore, S. (2017). New study finds face with tears of joy is world's most popular emoji. *Daily Mail,* Retrieved Jan. 30, 2017, from http://www.dailymail.co.uk/sciencetech/article-4089052/Crying-way-chart-Face-tears-joy-revealed-world-s-popular-emoji.html

[12] Dearden, L. (2016). Istanbul nightclub attack: At least 39 dead after gunman 'dressed as Santa' opens fire at new year's party. *The Independent - Europe, Independent,* Retrieved Jan. 30, 2017, from http://www.independent.co.uk/news/world/europe/istanbul-nightclub-attack-shooting-turkey-gunman-injured-killed-victims-new-years-eve-2016-2017-a7504046.html

[13] SentiWordNet. *SentiWordNet*, Retrieved Jan. 30, 2017, from http://sentiwordnet.isti.cnr.it/

Mohammed O. Shiha was born in Jerusalem, Palestine. He received the B.S. degree in computer engineering from Al-Quds University, Palestine in 2014. He is currently a master's degree student in the Department of Computer Engineering at Bahcesehir University, Istanbul, Turkey. He is working in the area of software development and computer forensics in Istanbul, Turkey.

Serkan Ayvaz received his bachelor’s degree in mathematics and computer science in 2006 from Bahçeşehir University Istanbul, Turkey. Later, he received his master’s degree in technology with specialization in computer technology from Kent State University in 2008. He completed his Ph.D. in computer science at Kent State University in 2015. He has over 8 years of industry work experience in the USA, most recently as a lead systems analyst at the Cleveland clinic foundation between 2011 and 2016. In his role, he served on multidisciplinary research teams focusing on medical research projects. Prior to joining the Cleveland Clinic, he had worked as a software engineer at Hartville Group for three years.

Ayvaz is currently a faculty member at the Department of Software Engineering and serves as the coordinator of the Big Data Analytics and Management Graduate Program at Bahcesehir University. His research interests include semantic searches, machine learning and scalable knowledge discovery in big data semantic web and its applications, particularly in healthcare and the life sciences.