*Proceedings of the 2nd Faculty of Industrial Technology International Congress*
*International Conference*
*Bandung, Indonesia, January 28-30, 2020*
*ISBN 978-623-7525-37-0*

# Clustering Emotional Features using Machine Learning in Public Opinion during the 2019 Presidential Candidate Debates in Indonesia

**Agus Sasmito Aribowo[1,2], Yuli Fauziah[1], Halizah Basiron[2], Nanna Suryana Herman[2], and Siti Khomsah[3]**

[1] Department of Information Engineering, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Yogyakarta - INDONESIA

[2] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka - MALAYSIA

[3] Department of Information Engineering, Institut Teknologi Telkom Purwokerto, Purwokerto - INDONESIA

Corresponding author e-mail: sasmito.skom@upnyk.ac.id, siti@itttelkom-pwt.ac.id

## Abstract

This research has produced a description of the emotions of streaming-video viewers of presidential candidate debates broadcasted on Youtube. In the first presidential candidate debate, the emotions of viewers were still neutral and tended to be feelings of pleasure and happiness. In the second to fifth presidential candidate debates, the dominant emotions were happy, angry, and sad. This research is known as emotion analysis, using comments from viewers of presidential candidate debates on Youtube as the data. Those comments were downloaded and pre-processed for data cleaning, emotion feature extraction, and clustering using K-Means based on six basic types of emotions: anger, sadness, happiness, fear, surprise, and disgust. The aims to be achieved are to determine a more homogeneous cluster for each opinion in the presidential candidate debate videos and to provide an emotional label for each cluster formed. The results of the research are five clusters that have distinctive homogeneity, namely happiness, anger, neutral, surprise-angry-disgust, and sadness. Each cluster member was labeled according to its characteristics. After being divided for each stage of the presidential candidate debate, it can be seen that the journey from the first debate to the next debate period tended to increase the emotion of anger and reduce the emotion of neutral.

*Keywords: Emotion Analysis, Clustering, K-Means, Basic Emotion*

## 1. Introduction

Indonesia entered the political year in 2019. Political parties make use of social media as a means of political communication by posting political news, comments and videos. Youtube as the largest video-based news portal is used as a means of political communication. Comments on Youtube videos related to the presidential election represent public response to the political figures, campaign programs, and political promises. Analysis of sentiment is needed to analyze and assess public opinion on a video, conversation, or certain topics on social media. Sentiment analysis is the extraction of information which aims to extract information about the feelings of authors both positively and negatively by analyzing many documents (Mukherjee & Bhattacharyya, 2013). Sentiment analysis can be considered as a combination of text mining and natural language processing. Basically, the work process of text mining adopts many data-mining studies, but the difference is that the patterns used by text mining are taken from a group of unstructured natural languages whereas in data mining, the pattern is taken from structured data (Han & Kamber, 2012). An important feature of analysis sentiment is the emotion feature. Emotion analysis is also a field of study that analyzes emotions towards entities such as products, organizations, individuals, information services, issues, events, topics, and attributes. The emotion analysis is computational research of emotions expressed textually (Liu, 2012). Many of English emotion analysis researches classify documents or opinions based on Ekman's 6 (six) basic emotions: surprise, happiness, anger, fear, disgust, and sadness (Ekman, 1992).

It is difficult to classify public opinion on Youtube videos, that are available in large numbers, into several groups of emotions. In addition, related research is still limited. Then, this research is about how to recognize emotions so as to expand the level of existing sentiments. The aim is to describe public emotions better so that it can be known at what level they are considered harmful.

One of the methods for unsupervised grouping of data is clustering method. This method can be used as a first step to group opinion data based on emotional features that have been extracted previously. Then, the most dominant characteristics in each cluster produced can be observed and labeled. The label will be used as a label for every data in the cluster. The results of the clustering will be visualized, and it functions as a description of the emotions of netizens in each presidential candidate debate.

## 2. Literature review

Research on sentiment and emotion analysis includes predicting the results of gubernatorial or presidential election (Mochamad Ibrahim, Abdillah, Wicaksono, & Adriani, 2015) (Joyce & Deng, 2017) (Budiono,

Nugroho, & Doewes, 2017) (Attarwala, Dimitrov , & Obeidi, 2017) (Kušen & Strembeck, 2018), predicting the results of parliamentary elections (Smailovic, Kranjc, Grcar, Znidarsic, & Mozetic, 2015) (Castro & Vaca, 2017), political party market share (Sharma & Moh, 2016) (Wicaksono, Suyoto, & Pranowo, 2016), political figures (Razzaq, Qamar, & Bilal, 2014) and general political conditions in countries (Charalampakis, Spathis, Kouslis, & Kermanidis, 2015) (Filho, Almeida, & Pappa, 2015). If emotion analysis is applied to public opinion on the 2019 presidential candidate debate video, it is expected that public emotions toward the content of the presidential candidate debate videos can be known. This certainly gives positive contributions to politics in Indonesia. Emotion features are needed to determine the power of sentiment and are divided into six basic emotions according to Ekman, namely anger, sadness, disgust, happiness, fear, and surprise. It is necessary to specify how negative an opinion is because of the fact speaking, especially negative sentiment. There are often incidents in social media. Social media is often used for negative political communication, such as black campaigns, negative sentiment, hate speech, and opinion warfare. Social media is often used for hate speech between supporters of political figures against others. Many incidents of political fanaticism on social media cause broken friendships, broken brotherhood, destroying family and community relations (BBC News, 2014). There are cases of political fanaticism on social media causing murders (Flo, 2018).

## 3. Methodology

### *3.1 Research Steps*

Based on the background of the problem, the research steps are shown in Fig 1. The research went through 6 (six) steps and was carried out sequentially. This sequence of steps were carried out to group political opinions in an unsupervised way on social media related to the presidential candidate debates to display the results in graphical form.

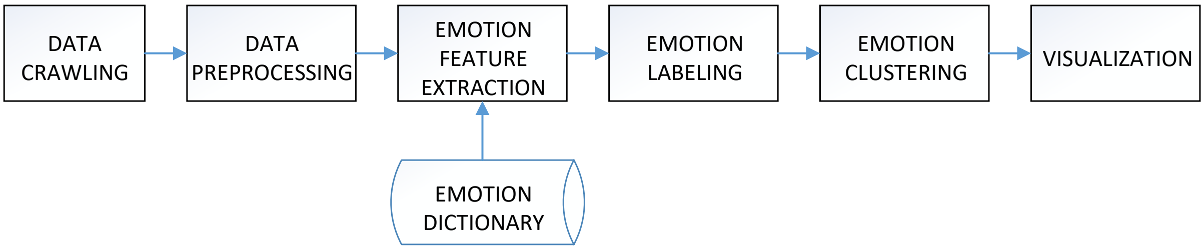


**Fig. 1: Research Steps**

### *3.2 Data Crawling*

Public opinion on the videos of the presidential candidate debates was obtained from the official video of the television stations that broadcasted the debates live using streaming media on Youtube. Not all channels were selected in the data collection. The debate schedule, topics, and Youtube channel that is the source of the data are in Table 1:

**Table 1: Debate Schedule and Youtube Channel**

| Debate Number | Topic | Channel | Record Number |
|---|---|---|---|
| I | Law, Human Rights, Corruption and Terrorism | Kompas TV<br>IDN Times | 715 |
| II | Energy and Food, Natural Resources and the Environment, and Infrastructure | MNC TV<br>CNN Indonesia | 715 |
| III | Education, Health, Employment, Social and Culture | Kompas TV<br>IDN Times | 715 |
| IV | Ideology, Governance, Defense and Security, International Relations | Kompas TV<br>IDN Times | 715 |
| V | Economy and Social Welfare, Finance and Investment, Trade and Industry | TV One News<br>CNN Indonesia | 715 |
|  | TOTAL |  | 3575 |

Two Youtube channels were chosen in every presidential candidate debate. Youtube channel selection is based on the largest number of viewers, video length, completeness of video content, and the highest number of comments. The data retrieval process was done with Youtube Comment Scrapper software, and all comments from the videos could be downloaded and saved in csv format. The csv file had a good structure, containing comments and comment replies, date, users' identities, and several other attributes. Selection of comments in csv file was done manually, so that only opinion comments related to the videos of presidential candidate debates would be included in the next process. There are 715 opinions chosen randomly in each debate.

## *3.3 Data Preprocessing*

Data preprocessing activities are removing unstructured text, converting text into words that are easily processed by machine learning, and deleting unimportant text data. Pre-processing is crucial in sentiment analysis because social media mostly contains unstructured words. The purpose of preprocessing is to clean and make the word uniform so that the word is ready for extraction to the next stage (Haddi, Liu, & Shi, 2013). There are several sub-tasks of pre-processing such as tokenizing, removing punctuation, removing username, removing hashtag, cleaning number, cleaning one character, removing URL, removing RT i.e. the symbol "@" before a username in question, converting non-standard words and emoticons. Preprocessing is useful for experts to facilitate the labeling of emotions so that preprocessing does not eliminate punctuation, word affixes, and letter case.

## *3.4 Emotion Feature Extraction*

After the opinions were cleaned in the data cleaning stage, the emotional features of the opinions were removed from each sentence. The process of extracting emotional features from each sentence was done by matching each word in the opinion sentence with the NRC-Emotion-Lexicon dictionary downloaded from the internet (Mohammad, 2017). This dictionary contains a list of keywords along with the dominant types of emotions that apply to those keywords. Each sentence can contain several emotional features, and each emotional feature can have different strengths. However, the emotional features produced by matching with this dictionary are not the final justification for each opinion sentence. This research involves experts who manually test every sentence feature generated using the dictionary.

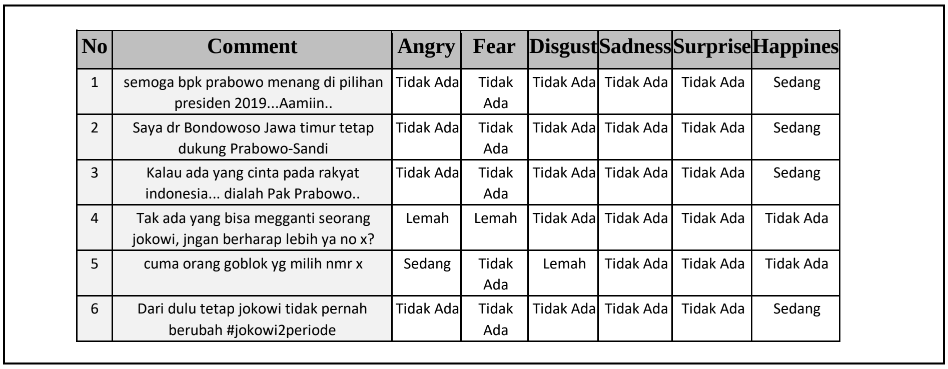
## *3.5 Emotion Labeling*

Emotion is divided into 6 types according to Ekman's 6 (six) basic emotions such as happiness, surprise, anger, disgust, fear, and sadness (Ekman, 1992). In this research, every type of emotion has 4 (four) levels of polarity as shown in Table 2.

**Table 2: Level and Weight of Emotion**

| Level of Emotion | Weight | Information |
|---|---|---|
| High | 3 | The emotion value on the text approximately 100%-67% |
| Average | 2 | The emotion value on the text approximately 66%-34% |
| Low | 1 | The emotion value on the text approximately 33%-1% |
| None | 0 | No emotion on the text |

Emotion labeling for each opinion sentence that has been cleaned and extracted using a dictionary was done by experts. Experts provided justification for emotional judgment in each sentence using the emotional assessment form as in Fig. 2

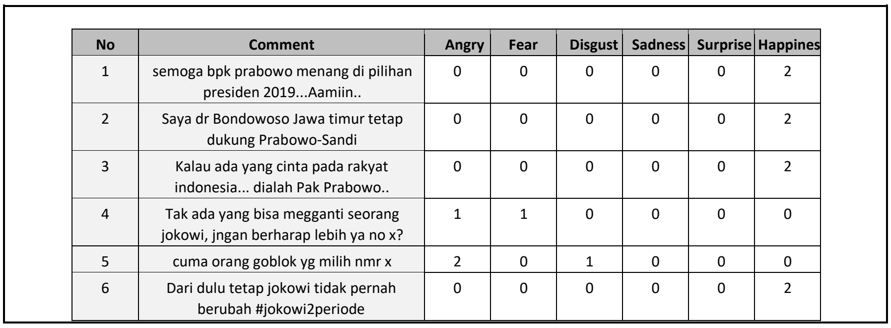


| No | Comment | Angry | Fear | Disgust | Sadness | Surprise | Happines |
|---|---|---|---|---|---|---|---|
| 1 | semoga bpk prabowo menang di pilihan presiden 2019...Aamiin.. | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Sedang |
| 2 | Saya dr Bondowoso Jawa timur tetap dukung Prabowo-Sandi | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Sedang |
| 3 | Kalau ada yang cinta pada rakyat indonesia... dialah Pak Prabowo.. | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Sedang |
| 4 | Tak ada yang bisa megganti seorang jokowi, jngan berharap lebih ya no x? | Lemah | Lemah | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada |
| 5 | cuma orang goblok yg milih nmr x | Sedang | Tidak Ada | Lemah | Tidak Ada | Tidak Ada | Tidak Ada |
| 6 | Dari dulu tetap jokowi tidak pernah berubah #jokowi2periode | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Tidak Ada | Sedang |

**Fig. 2: Form to determine the type of emotion**

### *3.6 Clustering*

The clustering process starts by converting each emotion label into a number. The conversion number is obtained from the weight in Table 2. For positive emotions, the weight is positive (happiness and surprise). For negative emotions, the label will be given a negative value (anger, sadness, fear and disgust). Some of the results of the conversion of emotion labels into emotion weight are in Fig. 3.



| No | Comment | Angry | Fear | Disgust | Sadness | Surprise | Happines |
|---|---|---|---|---|---|---|---|
| 1 | semoga bpk prabowo menang di pilihan presiden 2019...Aamiin.. | 0 | 0 | 0 | 0 | 0 | 2 |
| 2 | Saya dr Bondowoso Jawa timur tetap dukung Prabowo-Sandi | 0 | 0 | 0 | 0 | 0 | 2 |
| 3 | Kalau ada yang cinta pada rakyat indonesia... dialah Pak Prabowo.. | 0 | 0 | 0 | 0 | 0 | 2 |
| 4 | Tak ada yang bisa megganti seorang jokowi, jngan berharap lebih ya no x? | 1 | 1 | 0 | 0 | 0 | 0 |
| 5 | cuma orang goblok yg milih nmr x | 2 | 0 | 1 | 0 | 0 | 0 |
| 6 | Dari dulu tetap jokowi tidak pernah berubah #jokowi2periode | 0 | 0 | 0 | 0 | 0 | 2 |

**Fig. 3: Converting from an emotion label into an emotion weight**

The clustering process was carried out using K-Means method with a configuration of a K value of 5 because it was desired to produce 5 clusters. The results of the clustering process are a table containing the centroid values of the results of the clustering process in the opinion data in Fig 3. Each centroid cluster becomes the midpoint of more homogeneous emotion groups. The emotion values for each cluster center point are in Table 3. Based on Table 3, this research concludes that each cluster has specific emotion characteristics. Clustering also manages to divide data into 5 large groups that are more homogeneous. The numbers in the highlighted table cell are the most dominant type of emotion in cluster 1 is anger (-2,36). Cluster 2 is happiness (2,19). Cluster 3 is surprise (2,68), anger (-2,027), and sadness (-1,19). Cluster 4 is sadness (-2,14). Negative number in Table 3 indicates the type of negative emotion and positive numbers indicate the type of positive emotion. The summary characteristics of each cluster are in Table 4. The dominant emotion characteristics in each cluster are the label for all data bound to the cluster. Thus, there are 5 data groups.

**Table 3: Centroit Table**

| Emotion | | | Cluster | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Happiness | 0.29 | 0.09 | 2.19 | 0.33 | 0.07 |
| Disgust | -0.13 | -0.66 | -0.02 | -1.86 | -0.09 |
| Surprise | 0.03 | 0.06 | -0.02 | 2.68 | 0.06 |
| Anger | -0.23 | -2.36 | -0.06 | -2.03 | -0.24 |
| Sadness | -0.14 | -0.21 | -0.02 | -1.19 | -2.14 |
| Fear | -0.14 | -0.21 | -0.04 | -0.92 | -0.09 |

**Table 4: Emotion Characteristics of each Centroit**

| Cluster | Emotion Characteristics Domination of each Cluster |
|---|---|
| 0 | Neutral |
| 1 | Anger |
| 2 | Happiness |
| 3 | Surprise, Anger, and Disgust |
| 4 | Sadness |

## 3.7 Data Labeling

Each sentence in the dataset is labeled as a cluster number. Then each presidential debate is recapitulated based on the number of members of each cluster as in table 5.

**Table 5: The number and percentage of each cluster in each presidential debate**

| Debate I | | | Debate II | | | Debate III | | | Debate IV | | | Debate V | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clus ter | Rec. Num | % | Clus ter | Rec. Num | % | Clus ter | Rec. Num | % | Clus ter | Rec. Num | % | Clus ter | Rec. Num | % |
| 0 | 465 | 65,3 | 0 | 402 | 56,4 | 0 | 232 | 32,7 | 0 | 252 | 35,7 | 0 | 301 | 42,1 |
| 1 | 33 | 4,6 | 1 | 86 | 12,0 | 1 | 95 | 13,3 | 1 | 226 | 31,7 | 1 | 174 | 24,3 |
| 2 | 200 | 28,0 | 2 | 226 | 31,6 | 2 | 353 | 49,3 | 2 | 234 | 32,7 | 2 | 239 | 33,6 |
| 3 | 2 | 0,0 | 3 | 2 | 0,0 | 3 | 2 | 0,0 | 3 | 3 | 0,0 | 3 | 1 | 0,0 |
| 4 | 14 | 2,0 | 4 | 0 | 0,0 | 4 | 33 | 4,7 | 4 | 0 | 0,0 | 4 | 0 | 0,0 |

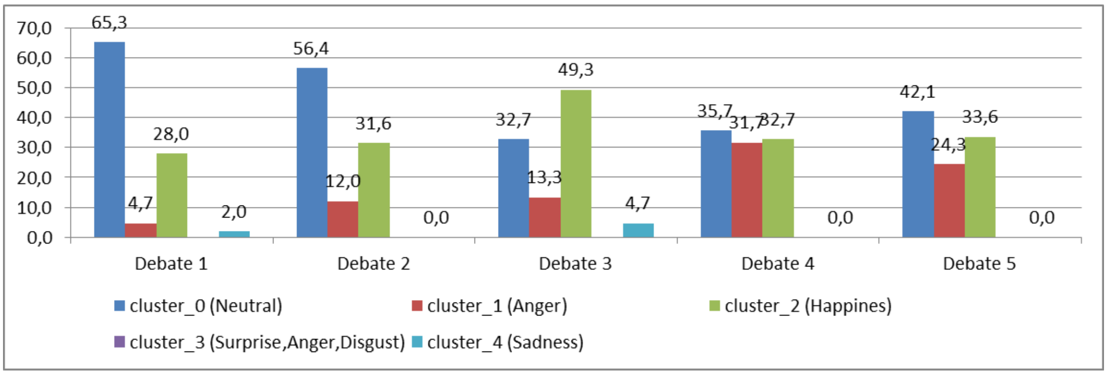


**Fig. 4: The percentage of clusters at each presidential candidate debate**

Fig 4 show than in the first presidential candidate debate, the dominant cluster was cluster 0, meaning that the audience was still neutral (65.3%) and did not show significant emotions. In the second presidential candidate debate, viewers began to show the emotion of anger (12%) and happiness (31.6%). In the third presidential candidate debate, viewers showed increasingly diverse emotions, ranging from anger (13.3%), happiness (49.3%) and sadness (4.7%). In the fourth debate, viewers showed only two kinds of emotions, namely anger (31.7%), and this was the peak of the emotion of anger, and happiness (32.7%). In the fifth presidential candidate debate, anger decreased (24.3%), and the emotion of happiness increased slightly (33.6%).

## 4. Conclusion

This research produces a description of emotions of viewers of presidential candidate debates on Youtube using K-Means machine learning methodology. The description of the emotions of the viewers shows that the most dominant types of emotions are happiness and anger. These emotions generally begin to emerge in the second presidential candidate debate. This description can be used to monitor interactions between viewers of presidential candidate debates. If negative emotions tend to strengthen or increase in frequency, this monitoring can be an early warning towards conditions that threaten national security.

## 5. References

Attarwala, A., Dimitrov, S., & Obeidi, A. (2017). How Efficient is Twitter - Predicting 2012 U.S. Presidential Elections using Support Vector machine. In *Intelligent Systems Conference 2017*.

BBC News. (2014). Putus Pertemanan Gara-Gara Pilpres. *BBC News*.

Budiono, D. F., Nugroho, A. S., & Doewes, A. (2017). Twitter Sentiment Analysis of DKI Jakarta's Gubernatorial Election 2017 with Predictive and Descriptive Approaches. In *International Conference on Computer, Control, Informatics and its Applications Twitter*.

Castro, R., & Vaca, C. (2017). National Leaders' Twitter Speech to Infer Political Leaning and Election Results in 2015 Venezuelan Parliamentary Elections. In *International Conference on Data Mining Workshops National*.

Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2015). Detecting Irony on Greek Political Tweets : A Text Mining Approach. *16th EANN Workshops*.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion, 6*(3–4), 169–200.

Filho, R. M., Almeida, J. M., & Pappa, G. L. (2015). Twitter Population Sample Bias and its impact on Predictive Outcomes. In *International Conference on Advances in Social Networks Analysis and Mining*.

Flo, E. (2018). Fanatisme Dukung Capres Berujung Pembunuhan, Ini Tanggapan Presiden Jokowi. *MerahPutih.Com*.

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *First International Conference on Information Technology and Quantitative Management*, *17*(December 2014), p.26–32.

Han, J., & Kamber, M. (2012). Data Mining: Concepts and Techniques Jiawei. In *Data Mining: Concepts and Techniques Jiawei*.

Joyce, B., & Deng, J. (2017). Sentiment Analysis of Tweets for the 2016 US Presidential Election. *IEEE*, 5–8. Kušen, E., & Strembeck, M. (2018). Politics, Sentiments , and Misinformation : An Analysis of the Twitter Discussion on The 2016 Austrian Presidential Elections. *Online Social Networks and Media 5, 5*, 37–50.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypoll Publisher.

Mochamad Ibrahim, Abdillah, O., Wicaksono, A. F., & Adriani, M. (2015). Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. In *15th International Conference on Data Mining Workshops Buzzer*.

Mohammad, S. M. (2017). NRC Word-Emotion Association Lexicon.

Mukherjee, S., & Bhattacharyya, P. (2013). *Sentiment Analysis : A Literature Survey*. *Indian Institute of Technology, Bombay*.

Razzaq, M. A., Qamar, A. M., & Bilal, H. S. M. (2014). Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*.

Sharma, P., & Moh, T.-S. (2016). Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter. In *IEEE International Conference on Big Data (Big Data) Prediction*.

Smailovic, J., Kranjc, J., Grcar, M., Znidarsic, M., & Mozetic, I. (2015). Monitoring the Twitter sentiment during the Bulgarian Elections. *IEEE*.

Wicaksono, A. J., Suyoto, & Pranowo. (2016). Proposed Method for Predicting US Presidential Election by Analysing Sentiment in Social Media.pdf. In *2nd International Conference on Science in Information Technology (ICSITech)*.