Topic Modeling and Sentiment Analysis of Martial Arts Learning Textual Feedback on YouTube

John Paul P. Miranda[1], Jonar T. Martin[2]
[1]Don Honorio Ventura State University, Philippines, johnpaul.p.miranda@gmail.com
[2]Don Honorio Ventura State University, Philippines, jonarmartin@gmail.com

ABSTRACT

Learning through videos preferably with the use of YouTube indicates a shift in gaining knowledge where the role of visual learning is essential to a new generation of learners. Studies confirmed that YouTube was able to maintain the interest and learning accomplishments of the students. In this study, YouTube textual feedback dataset containing keywords related to the learning of martial arts is examined using topic modeling and sentiment analysis. Topic modeling using Latent Dirichlet Allocation (LDA) was implemented to identify the topics within the first corpus, and then the NLTK library were used to determine the sentiment classification of the corpus through Jupyter Notebook. Sentiment analysis indicated that generally the discussion on martial arts is positive, more so when the viewers are signifying their intention to learn and reacting to the impressive execution of the martial arts. Topic modeling indicated that the numerous topics of discussion on martial arts are closely related to learning, arts, and humanities. Results can help educators, martial arts trainers, developers of YouTube content, and educational institutions improve resources or trainings that might help them achieve educational or physical performance targets for learners.

Key words : Martial arts, sentiment analysis, YouTube, text mining, topic modeling.

1. INTRODUCTION

The internet has changed our lives forever on how we do business, how we communicate with other people and how we learn. According to Siegel [12], the use of technology featuring videos in education particularly inside the classroom is promising specifically towards the 21st Century Learning Framework. In a great degree, learners can now easily connect to different learning platforms to enhance their learning experience, one of which is through instructional videos. Videos can be easily accessed through the World Wide Web, with YouTube as the well-known platforms for disseminating video content where interest on its educational use has become apparent [7], [23]. In a survey by global market research firm, YouTube was listed as the top choice platform for learning by the so-called Generation Z (i.e., people ages 14 to 23) and it was revealed that 59% preferred YouTube as a learning resource [20].

YouTube preference or video learning signal a change in learning styles where the role of visual learning is crucial to a new generation of learners. People turn to YouTube now due to its accessibility and its content as well as it powers the users to easily control the pace of the video which allows them to absorb the content at their own pace. [20]. The study of Orús et al. [6] confirmed YouTube's positive effect on learning outcomes and student satisfaction. Similarly, Chtouki et al. [26] found out in their study students are better exposed in visual explanation video in understanding and remembering complex concepts in which is one of the main advantages of YouTube. It was also concluded by Azer et. al, [27] that a small number of YouTube videos can be useful towards independent learning.

YouTube's educational value has been found consistently in a number of disciplines. In math competency, Fulton [18] determined that YouTube enhances learning in mathematics as evidenced by competency scores. In the field of nurse education, Cliffton and Mann [1] discovered that using YouTube videos enhances learning, engagement, and critical awareness among students. Similarly, Ramos-Rincón et. al [28] pointed out that by producing YouTube videos, it can contribute to the learning processes by promoting personal responsibility while reinforcing the skills of the students and maintaining their interest. Meanwhile, in a study by Tan and Pearce [13] revealed that YouTube videos can be attributed towards supporting students' understanding and an effective way to engage them. However, students in the study still perceived the role and authority of the teachers which are hard to ignore after introducing the video resources.

The use of YouTube has also benefited those who seek learning when it comes to the physical performance where they can view the execution of the motor technique

repeatedly. In connection to this, Dewitt et al. [9], reported that YouTube was a valuable teaching resource in sustaining students' interest and learning achievement in the areas capturing current performing arts trends such as dance, animation, and fine arts. Another discipline, which gained traction when it comes to subscription, are videos on martial arts. Martial arts practitioners claim that they learned difficult martial arts move through only watching YouTube [22]. However, yet to date, literature is limited on how forms of media like films, documentaries, YouTube videos, images, and blogs might shape the actual "hands-on," solo and physical training of the martial arts. Further research on educational evaluation of YouTube videos and the resources offered by the video sharing service is needed [7].

One function of YouTube is the input or comment section where viewers write their thoughts or sentiments, in which students react to the content of resources. Zher Ng and Maznah Raja Hussain [16], [17] emphasized that understanding the textual feedback is critical as it ensures learning happens. According to Dahal, Kumar, and Lee [4], said that topic modeling and sentiment analysis can be used as an approach for this context. Blei [8] defined topic modeling as a technique of finding a secret semantic structure in a corpus (i.e. a large or full set of writings) providing insights into the various themes in the texts. Meanwhile, sentiment analysis is the method of defining the emotions and thoughts expressed in a given text [25]. Furthermore, it is an emerging field for classifying phenomenon found in text data [15], [17]. It is also a form of text mining technique which involves the calculation of textual data obtained from user-generated contents from different platforms to identify its positivity, negativity, and neutrality [5], [24].

Recent studies have shown that using these techniques of text mining offered great insights on finding values on products and services [29] and also into the types of discussion that can be found on YouTube or similar platform like Twitter [3-4] [21]. Sailunaz and Alhajj [19] used the tweets from twitter to analyze the sentiments conveyed from users. Furthermore, the discovered results from specific topics were used to provide recommendations to its users. The research aims to add to the discussion threads about the learning context of YouTube videos. Analyzing the learners' feedback and sentiments will help educators, martial arts teachers, YouTube content creators, and educational organizations enhance tools or preparation that can help them meet educational or physical performance goals for learners. The study aims to describe the feedback of YouTube users in learning Martial Arts through topic modeling and sentiment analysis in terms of the number of feedbacks, most occurring words, its classifications and sentiments. Furthermore, this study explores how the inferred topics and sentiments from the feedbacks can be used to create an instructional video for martial arts learning and physical performance in general.

2. METHODS

The study aims to perform sentiment analysis and topic modeling through an extracted YouTube feedbacks which will be the dataset of this study. Several steps were performed sequentially to the dataset, such as data collection, data cleaning, data preparation [4], and topic discovery [2].

2.1 Data Extraction Process, Sample Size and Sampling Design

Feedbacks from Martial Arts video tutorials from YouTube were extracted, pre-processed, and visualized in Jupyter Notebook, particularly Python programming. The "Martial Arts" keyword was used to search for the videos. Additional parameters were included to the YouTube Data Application Programming Interface (API) such as sorting the video results by its relevance to the keyword and to English language. The API was able to return 1 million videos available, but only 950 videos can be accessed due to the country restriction. Afterward, a total of 610 videos were pre-selected after removing the videos with disabled comments or doesn't contain any comments. Subsequently, videos were further filtered using their video titles, and description, only videos containing the keywords such as "instruction," "tutorial," "learn," "teach," "basic," "beginner," "practical," and "training" was selected. Finally, only the top 100 feedbacks ordered by its relevance in each video were selected to form the final dataset used in the study. The dataset, which consists of 20, 586 feedbacks, were then stored into a spreadsheet.

2.2 Data Cleaning and Data Pre-processing

For data preparation towards sentiment analysis and topic modeling, this pre-processing and cleaning is implemented to the dataset in order to reduce the noise within the dataset, its dimensionality and enhance the effectiveness of sentiment analysis and topic modeling. In data cleaning, all unrecognizable characters, non-English comments, all emails, special characters such as new line characters, quotes, and other web links were removed. Subsequently, in data pre-processing, the dataset was exposed to additional steps. First, the removal of numbers, stemming, part of speech tagging, converting to lower case, and finally using the Natural Language Toolkit (NLTK) library for filtering and excluding stop words. Also, the dataset was tokenized and subjected to the Gensim's Phrases Model to produce the bigrams and trigrams of frequently occurring phrases within the dataset to form the first corpus. Second, the corpus was further filtered using Loughran-McDonald Sentiment Word Lists to form the second corpus.

2.3 Data Analysis

The corpora were analyzed using sentiment analysis and topic modeling. LDA was used as the model for obtaining the topic within text. It was implemented to determine the topics within the first corpus, and then the NLTK library were used to determine the sentiment classification of the corpus through Jupyter Notebook. In sentiment classification, two steps were used. Firstly, the sentiment intensity analyzer method from the NLTK library was applied to each feedback within the corpus to determine the overall sentiment. Secondly, the words within the corpus were classified into positive, negative, and uncertainty and then showed using a word cloud. The LDA was also used in the second corpus to find the four (4) clusters and presented them using word clouds.

For the topic modeling, the previous studies proposed to use topic coherence as an evaluation tool for topic models [10], [11]. The researchers used topic coherence for the number of optimal topics, this is executed by extracting the average coherence score per topic and selected the most important and interpretable topics by using the same measure and through averaging the coherence score in each topic and obtaining the one that has good coherence score.

3. RESULTS

The sentiment intensity analyzer method from the NLTK library was applied to determine the overall sentiment to each feedback within the corpus. As shown in Table 1, the top most occurring words are "good," which obtained the highest frequency, while "great" and "best" obtained almost the same frequencies. The word "good" signifies the intention to learn the skills in martial arts without the purpose to use it for offensive purpose (e.g. I just wanted to learn how to punch good not kill somebody). The feedback that contains the word "great" indicates the appreciation of the viewers on the effort in producing the video in terms of physical effort and relaying it with clarity (e.g., Great work, efforts are really important behind this video 'Gurukulam communication' for the traditional martial art, in situ). In contrast, from the extracted feedback, the words "could" and "maybe" are used to suggest that important information on the martial art be included such as the name and the style (e.g., Would be awesome if someone could Name all the fighting Styles used, in situ). It is also suggested the components of the video that would make it more appealing to the viewers (e.g., good video maybe we could do without the music next time, was kind of annoying lol, in situ).

The feedback "you r the best bro, in situ" which belong to the feedback "best" indicates commendation to the performer in the skillful performance of the martial art. While the word "better" indicates a comparison of execution of the person in the video to a popular personality in martial arts (e.g., Tony

has better techniques than donnie yen, in situ). The word "might" is used by users to convey their interest in other topics that the content creator might want to create next. The comment such as "I really want to start training more grappling, so mma might be the next step" demonstrates this finding. Meanwhile, the words "effective," "believe," and "wrong." are signs that the viewer's doubt the efficacy of the videos in terms of accuracy of the information in terms of applicability, authenticity, and epistemology.

Results indicated the intention of the viewers to learn and at the same time are interested in how the videos are designed. Also, appreciation among the viewers in terms of the execution is evident and they also provide suggestions on what is a more appealing video presentation. In contrast, the viewers warranted the sentiments on correct information.

Table 1: Top 10 Most Occurring Words in Martial Arts YouTube Videos

Rank	Words	Frequency	Sample Feedback
1	good	1353	I just wanted to learn how to punch good not kill somebody
2	great	1091	Great work, efforts are really important behind this video 'Gurukulam communication ' for the traditional martial art.
3	best	1080	you r the best bro
4	could	409	Would be awesome if someone could Name all the fighting Styles used :DD
5	better	386	Tony has better techniques than donnie yen
6	effective	260	It's almost like traditional martial arts moves are effective if, AND ONLY IF, you learn how to apply and integrate them into actual fighting scenarios
7	believe	187	I believe in energy, but these videos are completely counterfeit. But

			this is not impossible.
8	might	137	I really want to start training more grappling, so mma might be the next step.
9	wrong	136	Aikido is based on the leading techniques of baguazhang and infused with Aikijutisu, so you're objectively wrong.
10	maybe	136	good video maybe we could do without the music next time, was kind of annoying lol

The words within the corpus are shown in Fig. 1 using word cloud which is classified into positive, negative, and uncertainty. The dominant positive word is "good." For example, one feedback said, "He did pretty good needs to diversify his trick set but his speed was amazing and work on his left to right arm tricks rather than relying on his right to left arm tricks otherwise I was impressed, in situ." It means that some of the viewers already have prior knowledge or experience in the martial arts. Having experience in performing the techniques themselves would be instrumental in their analysis of the appropriate movement.

Meanwhile, the prominent negative sentiment is the word "wrong" where one feedback said, "I can respect this video but truth be told you know very little about martial arts, in situ." Generally, the negative sentiments signify the wrong information on the martial arts. This implies that in creating a video presentation, a thorough review of the literature on the martial arts should be done to make it more reputable. The frequently used feedback for the uncertainty category is the word "could" which is a mixture of different suggestions on various topics that they have experienced in real life and connects them to the video that they watched.
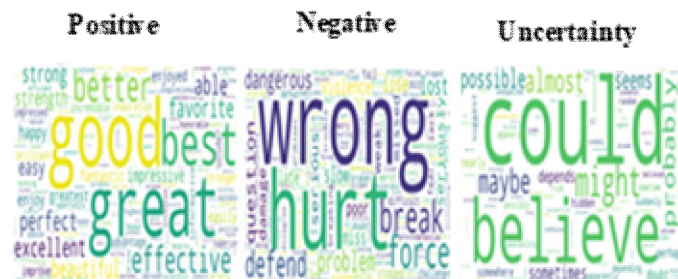

Figure 1: Sentiment word clouds generated from the corpus using Loughran and McDonald Word List

Table 2 displays topic modeling results on the feedback on YouTube about Martial Arts. Fig. 2 displays the word clouds of using the same results of the model. The top 10 most occurring words per topic is listed and the labels were added based on those words, the labels were chosen through the help of a martial arts expert and through examining the feedbacks to see what are those words is all about. The topics which emerged are the word label learning, art, competition, and self-defense. Topic modeling revealed that the discussions of martial arts are focused on learning and its importance to our life events.

Table 2: Result of Topic Modeling on the Feedback on YouTube Videos about Martial Arts

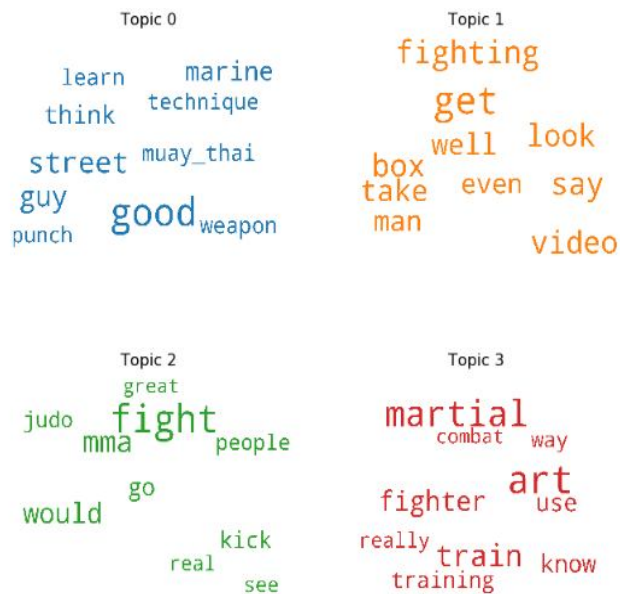| Topic	Label	Top 10 words
1	Learning	"good", "street", "guy", "marine", "think,", "weapon", "learn", "muay_thai", "technique", "punch"
2	Art	"get", "fighting", "box", "say", "video", "look", "take", "well", "man", "even"
3	Competition	"fight", "mma", "would", "go", "kick", "judo", "people", "see", "real", "great"
4	Self-defense	"art", "martial", "train", "fighter", "know", "use", "training", "way", "really", "combat"

Figure 2: Final topics found using LDA

Fig. 3 shows the topic coherence average in terms of the total volume of feedbacks. The mean was taken to facilitate direct comparison between the different feedbacks. The plot shows that the coherence score increases with the number of topics, with a decline between 4 to 5. According to Kumar (2018), in LDA it is advisable to build multiple models and choose the highest coherence score which is usually provides valuable

insights and easy to interpret. Hence, smaller subtopics are often seen when choosing a relatively higher value of coherence score. The result of the topic coherence analysis was the basis in deciding the number of topics which is 4.
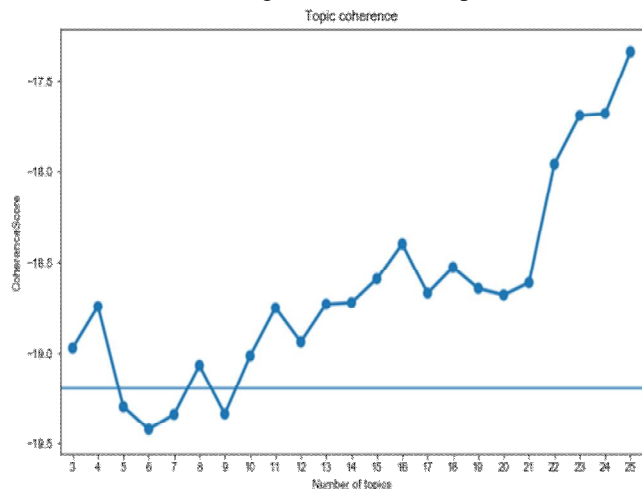

Figure 3: The optimal of topics using topic coherence is between 4 to 5.

Table 4 and Fig. 5 presents the polarity of words while Fig. 4 displayed the sorted results of tokenized words and color coded by their sentiment analysis in terms of contribution. Generally, martial arts learning videos received positive sentiments. Positive sentiments received 54% (n=6,733), Negative sentiments received 31% (n=3,986) and uncertainty sentiments received 15% (n=1,848). The words "good" (n=1,353), "great" (n=1,091) and "best" (n=1,080), respectively obtained the highest frequencies. Next to the positive sentiments which received high scores are uncertainty words, such as "could" (n=409) and "believe" (n=187). Negative sentiments have obtained the lowest scores such as the words "wrong" (n=136), "hurt" (n=109) and "break" (n=105). However, some of the frequencies are not shown in the figure since only sentiments with a score of 100 are included in the figure, but in totality, negative sentiments are more frequent that uncertain words. Findings indicated that the majority of the sentiments on martial arts is generally positive.
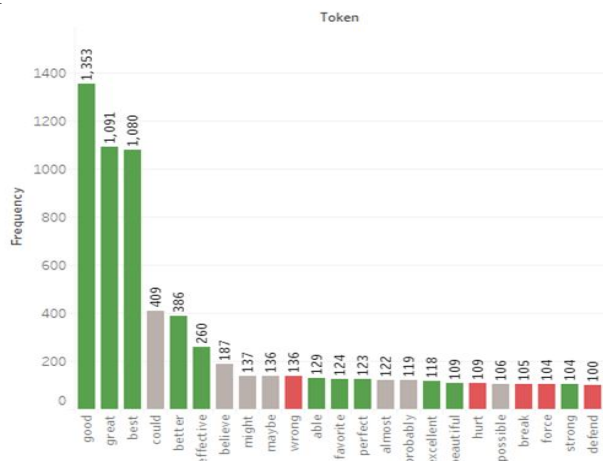

Figure 4: Sorted Martial Arts Video Feedbacks on YouTube Categorized by Sentiment Analysis

Table 4: Distribution of Words

Polarity	Percentage
Positive	6,773 (54 %)
Negative	3,986 (31 %)
Uncertainty	1,848 (15 %)
TOTAL	12, 607

4. CONCLUSION

This study examined the YouTube users' textual feedback in martial arts learning. It focused on analyzing their opinions through topic modeling and sentiments analyses. Lexicon-based sentiment analysis shows positive, especially when users are signifying their intention to learn and reacting to the impressive execution of the martial arts and it shows that learners tend to be expressive and appreciative towards the learning content of the videos which suited their needs. The learners also commended the efforts in producing quality videos and suggested elements that would make it more appealing to the viewers. Furthermore, precise execution and accurate information about the martial arts are warranted by the viewers. Topic modeling showed the clusters are related to learning, arts, and humanities. In particular, the discussions of martial arts are focused on learning and its importance to our life events.

This paper elaborated the positive responses of users towards YouTube martial arts learning content. This confirms that videos are valuable resources in learning discipline. In general, YouTube users elicited positive sentiments. It also confirms that martial arts is perceived to be an important aspect of our life events as people.

Instructional video producers can benefit from the findings in designing their martial arts video content. The sentiments and emerged topics will serve as a basis in creating an instructional video for learners of martial arts, and other disciplines. The suitability of the teaching method in the video also plays a great role in achieving its objective in teaching proper body mechanics in its execution. Since video production is also being used as a classroom learning activity, this will guide the teachers and the learners on the important aspects to consider in creating an attractive, informative, and learner-friendly learning material.

The study was limited by the viewer's liberty to write everything that they want. Therefore, there was no focus on the instructional component or teaching-learning process of the videos. It is suggested that video developers provide guidelines on how to give feedback that focuses on the teaching-learning aspect. Although this study presents how video content can be helpful towards martial arts learning, effectiveness of these videos towards in-class learning can be

explored. Also, a research in the same context or in other discipline but related to other learning sites and platforms may be worth pursuing. Future studies may be replicated featuring other disciplines or topics employing a mixed method or qualitative research design.

ACKNOWLEDGEMENT

REFERENCES

1. A. Clifton and C. Mann. **Can YouTube enhance student nurse learning?**, *Nurse Education Today*, vol. 31, no. 4, pp. 311-313, May 2011. https://doi.org/10.1016/j.nedt.2010.10.004

2. A. Karami, A. A. Dahlb, G. Turner-McGrievyb, H. Kharrazic, and G. Shaw Jr. **Characterizing diabetes, diet, exercise, and obesity comments on Twitter**, *International Journal of Information Management*, vol. 38, no. 1, pp. 1-6, Feb. 2018. https://doi.org/10.1016/j.ijinfomgt.2017.08.002

3. A. Steinskog, J. Therkelsen, and B. Gambäck, **Twitter topic modeling by tweet aggregation**, in *Proc. of the 21st Nordic Conference of Computational Linguistics*, Sweden, 2017, pp. 77-86. https://doi.org/10.1007/s13278-019-0568-8

4. B. Dahal, S. A. P. Kumar, and Z. Li. **Topic modeling and sentiment analysis of global climate change tweets**, *Social Network Analysis and Mining*, vol. 9 no. 24, pp. 1-20, June 2019.

5. B. Liu. **Sentiment Analysis and Opinion Mining**, *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp.1-167, May 2012.

6. C. Orús, M. J. Barlés, D. Belanche, L. V. C. Ariño, E. Fraj-Andrés, and R. Gurrea. **The effects of learner-generated videos for YouTube on learning outcomes and satisfaction**, *Computers & Education*, vol. 95, no. 1, pp. 254-269, Jan. 2016.

7. C. Snelson. **YouTube across the Disciplines: A Review of the Literature**, *MERLOT Journal of Online Learning and Teaching*, vol. 7, no. 1, pp. 159-169 March 2011.

8. D. Blei. **Probabilistic topic models**, *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, April 2012. https://doi.org/10.1145/2133806.2133826

9. D. DeWitt, N. Alias, S. Siraj, M. Y. Yaakub, J. Ayob, and R. Ishak. **The potential of Youtube for teaching and learning in the performing arts**, *Procedia - Social and Behavioral Sciences*, vol. 103, pp. 1118-1126, Nov. 2013.

10. D. Newman, J. H. Lau, Karl Grieser, and T. Baldwin, **Automatic evaluation of topic coherence**, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*, USA, 2010, pp. 100-108.

11. D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. **Evaluating topic models for digital libraries**. in *JCDL '10: Proc.s of the 10th annual joint conference on Digital libraries*, Australia, 2010, pp. 215-224. https://doi.org/10.1145/1816123.1816156

12. D. Siegle. **Technology: Literacy in the 21st Century: The Fourth R—Video Recording**. *Gifted Child Today*, vol. 32, no. 2, pp. 14-19, Jan. 2009.

13. E. Tan and N. Pearce. **Open education videos in the classroom: exploring the opportunities and barriers to the use of YouTube in teaching introductory sociology**, *Research in Learning Technology*, vol. 19, pp. 125-133. Jan. 2012. https://doi.org/10.3402/rlt.v19s1/7783

14. F. Koto and M. Adriani. **A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?**, in *Natural Language Processing and Information Systems*, Germany, vol. 9103, 2015, pp. 453-457.

15. H. Rahab, A. Zitouni, and M. Djoudi. **SANA: Sentiment analysis on newspapers comments in Algeria**, *Journal of King Saud University - Computer and Information Sciences*, pp. 1-9, April 2019. https://doi.org/10.1016/j.jksuci.2019.04.012

16. H. Z. Ng and R. M. R. Hussain. **Empowering learners as the owners of feedback while YouTube-ing**, *Interactive Technology and Smart Education*, vol. 6, no. 4, pp. 274-285, Nov. 2009

17. J. Ye, X. Jing, and J. Li. **Sentiment Analysis Using Modified LDA**, in *Signal and Information Processing, Networking and Computers, Lecture Notes in Electrical Engineering*, China, vol. 473, 2017, pp. 205-212.

18. K. Fulton. **The Flipped Classroom: Transforming Education at Byron High School: A Minnesota High School with Severe Budget Constraints Enlisted YouTube in Its Successful Effort to Boost Math Competency Scores**, *T H E Journal (Technological Horizons In Education)*, vol. 39, no. 3, pp. 18-20, April 2012.

19. K. Sailunaz and R. Alhajj. **Emotion and sentiment analysis from Twitter text**, *Journal of Computational Science*, vol. 36, Sept. 2019. https://doi.org/10.1016/j.jocs.2019.05.009

20. L. Genota. **Why Generation Z Learners Prefer YouTube Lessons Over Printed Books**, *Education Week*, Sept. 2018.

21. R. P. Bringula, J. N. Victorino, M. M. De Leon, and M. Regina Estuar. **Cluster and Sentiment Analyses of YouTube Textual Feedback of Programming Language Learners to Enhance Learning in Programming**, in *Proc. of the Future Technologies Conference (FTC) 2019*, USA, vol. 1070, 2020, pp. 913-924, Oct. 2019.

22. S. Kesting. **Can You Actually Learn to Grapple from Youtube and Video Games?**, *Grapplearts*, Feb. 2012.

23. T. Schirner. **The Advantages of YouTube in Education**, *Techwalla.*

24. V. Subramaniyaswamy, R. Logesh, M. Abejith, Sunil Umasankar, and A. Umamakeswari. **Sentiment Analysis of Tweets for Estimating Criticality and Security of Events**, *Journal of Organizational and End User Computing (JOEUC)*, vol. 29, no. 4, pp. 51-71, Dec. 2017.
https://doi.org/10.4018/JOEUC.2017100103

25. W. Medhat, A. Hassan, and H. Korashy, **Sentiment analysis algorithms and applications: A survey**, *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, Dec. 2014.

26. Y. Chtouki, H. Harroud, M. Khalidi and S. Bennani. **The impact of YouTube videos on the student's learning**, *2012 International Conference on Information Technology Based Higher Education and Training (ITHET)*, Istanbul, 2012, pp. 1-4.

27. S. Azer, H. AlGrain, R. AlKhelaif, and S. ALEshaiwi. **Evaluation of the Educational Value of YouTube Videos About Physical Examination of the Cardiovascular and Respiratory Systems,** *Journal of Medical Internet Research,* vol. 15, no. 11, pp. 1-13, Nov. 2013.

28. J. Ramos-Rincón, I. Belinchón-Romero, F. Sánchez-Ferrer, G. Martínez-de la Torre, M. Harris, and J. Sánchez-Fernández. **The reach of Spanish-language YouTube videos on physical examinations made by undergraduate medical students,** *Journal of Educational Evaluation for Health Professions,* vol. 14, pp. 1-8, Dec. 2017.
https://doi.org/10.3352/jeehp.2017.14.31

29. I. Juanatas, R. Fajardo, E. Manansala, A. Pasilan, J. Tabor, and H. D. Balmeo. **Sentiment Analysis Platform of Customer Product Reviews**, In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, pp. 230-234, Dec. 2019.
https://doi.org/10.1109/ICCIKE47802.2019.9004415