JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

VIDANALYTICS: ML-BASED VIDEO ANALYSIS FOR IDENTIFICATION OF TARGET AUDIENCE IN BUSINESS DEVELOPMENT

Dhruv Mehta
Dept. of Computer Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India mehtadhruv933@gmail.com

Jinay Parekh
Dept. of Computer Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India jinayparekh2001@gmail.com

Pratik Kanani
Dept. of Computer Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India pratikkanani123@gmail.com

Harvy Gandhi
Dept. of Computer Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India harvygandhi2@gmail.com

Tanya Sharma
Dept. of Electronics Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India tanyasharma130599@gmail.com

Aniket Kore
Dept. of Computer Engineering Dwarkadas. J. Sanghvi College of Engineering
Mumbai, India aniketkore24@gmail.com

Abstract—Content creators throughout the globe are con- fronted with an issue that cannot be solved by creativity. In today's competitive marketing landscape, ensuring that your product is advertised to the correct audience in the proper direction is critical. A video Indexer combines multiple audio and video Artificial Intelligence (AI) technologies into a single integrated service, simplifying development. This paper presents a clean, one-stop solution that will instantly assist anyone wishing to sell their product by abstracting away all the dirty API calls and loading windows. This is done by offering three core services: label identification (video analysis), keyword identification (audio analysis), and sentiment analysis. Combining insights from these channels provides a list of Reddit communities that the user can then target to market their content effectively. Furthermore, by recognizing the sentiments inside the video, the user has an advantage by acquiring specific target markets for where to distribute their content. Additionally, intelligent emotion filtering allows users to customize the message's tone.

Submitted: 17/10/2022 **Accepted**: 10/11/2022

Index Terms—LDA, SpaCy, Sentiment Analysis, NLTK Vader, Deep Face

I. INTRODUCTION

Digital technology has drastically changed the way businesses operate today. Companies have turned to digital marketing as a logical response to take advantage of and profit from the increased consumer focus on the Internet. The effects of digital marketing on people's interactions, routines, and lifestyles are overwhelming. After the onset of the pandemic, more and more people are connected to the internet and are logged onto social media. According to the Digital 2022: July Global Statshot report, 4.70 billion social media users around the world were recorded in July 2022, equating to 59.0 percent of the total global population. Approximately 93 percent of all website traffic comes from search engines. Before deciding to make a purchase, 82 percent of individuals use their phones to conduct an internet search. If a brand's name shows more than once during a search, 50 percent of consumers are more likely to choose it. With such a strong influence of a product's digital presence on consumer behavior, there is a need for efficient algorithms to map the desired product to the most accurate consumer preferences and patterns. The authors of [1] states that a new digital marketing environment has recently emerged as a result of the rapid growth of information and communication technology in both the public and private sectors. Data is currently generated in huge quantities as a result of the widespread adoption of information technology. [2] demonstrates the application of ML can help in making sense of vast amounts of data in a more time efficient and accurate manner. This method uses data to quickly spot patterns and anticipate future choices. Following that, marketers may utilize this data to optimize a significant chunk of their workflow, including increasing the number of tests they perform, enhancing the user experience of their website, and automating consumer engagement. Leaders hold a 97 percent consensus that digital marketers using Machine Learning-based solutions will shape marketing in the future. Content marketers need to have a deeper understanding of Machine Learning (ML) applications as complicated data becomes the standard. Unstructured data impedes performance and user experience because it is dispersed across platforms and takes many different forms. Using automated classification can help with this. The authors in [3] in order to categorize online news articles automatically, compare three leading ML methods: Random Forest, Neural Networks, K-Nearest Neighbors. The model presented intends to offer a simple, one-stop service that would be of immediate assistance to anyone looking to sell their product. The model captures discernible information from the marketing video, analyzes it and then outputs key data points associated with the video. It offers three key services to accomplish this: Sentiment Analysis, Label Identification (Video Analysis), and Keyword Identification (Audio Analysis). Based on the outputs of these services, it then scrapes a web page to find targeted audiences to the marketer. Thus, this model offers a one stop solution to video analytics.

II. RELATED WORKS

The tremendous implications of AI and ML on several sec- tors of the population remain to be felt significantly in the field of marketing. Despite this constraint, ML offers a number of important advantages, such as the possibility to utilise greater rigorous methodologies for generalising scientific discoveries.

[4] addresses this need by delving into marketing with a look at ML, covering its key types and algorithms, as well as its relevance to advertising and overall performance.

ML and AI have enormous promise for improving the knowledge and efficiency of advertising. In [5], a research evaluation of scientific journal publications on ML in diverse applications is conducted, and a proposed model detailing the key ML technologies and tools is supplied as the foundation for ML in marketing. To study such uses from 140 related papers, the 7Ps marketing plan is applied.

In [6] a smart maintenance decision support system for administering company information is explored. The analysis provided some insights into big data analytics and the decision-making process. The proposed research evaluated challenges in industrial big scale statistics utilizing a cost reduction technique. Using its analytical decision - making process, research work achieves superior predicted findings for case studies. Expertise and understanding are critical in reducing defects in production systems.

Using ML, [7] proposes a strategy for identifying people's requirements and managing their browsing potential. Based on the requirements of the person search criteria, the K-NN algorithm and other ML techniques in this process is utilized. This would result in a profitable business for individuals from all over the world by simplifying their everyday lives depending on their interests and hobbies. This strategy should be a categorized infrastructure through which can entice users to utilize the tool.

In [8], in order to improve image processing on movies via a cloud streaming platform, the research suggests SaW (Social at Work), a wireless network that offers a clear Web-based way. To lessen the overhead of computing on service systems, Grid Computing solutions leverage wireless technology. A networking site operator may save money by outsourcing video analysis to grid computing components as an architectural parallel to that of a flexible cloud storage service. Users can browse videos while client-side hardware is running batch image analysis tasks in the background while waiting for the server to combine the results.

On Reddit, a social media platform geared toward com- munities, users can post questions and express their opinions within subreddit communities and contribute to other users' discussions, rank, and comment. Social media platforms rely on recommender systems to steer interactions, and Reddit's recommender system operates at several levels. This research of [9] looks at the possible benefits of social media analytics for increasing recommendation performance. With an aim to enhance recommendations from subreddits that might potentially useful to a given user, five models are proposed and verified. The findings are consistent with the idea that in order to solve problems, it is essential to gather and combine a wide variety of attributes.

In [10], a dataset comprising words indicating user sentiments for various comments they have made on postings is examined. Reddit as the information source for news and comments is used to analyze the prediction of sentiment divergence to reactions to news items. Insights

into user behavior are acquired, comprehending the dynamics of interaction between users and their perspectives, and the bias towards certain topics on this platform through this study.

III. RESEARCH GAPS

Some works have aimed at using ML for marketing strategies. While some models have tried to use ML to find target audience for better marketing; they have done so by considering either of audio or video and not both. Numerous well- known annotation tools, including ViPER [11], LabelMe [12], and others, have been used in video object recognition and related activities. The majority of these labelling methods are interactive and need the involvement of a person. Key frame identification techniques in the past used a segmentation- based pipeline and these techniques typically extract SIFT and optical flow features [13]. Later research enhanced this field by employing keypoint detection for feature extraction [14]. The inherent flaw in all of these methods is that they could extract unnecessary important frames rather than properly encapsulating the video content. Moreover, there have not been great attempts at providing the user with targeted social media communities after analysing marketing videos. None of the prior works have been able to provide a one-stop service (i.e. analyse both audio and video, and provide a target audience at the end) that would be of immediate assistance to anyone looking to sell their product.

IV. METHODOLOGY

In this paper, the approach Figure 1 is divided into three stages: The first section discusses video analysis, which is used for label extraction, and the second section explains audio analysis, which is used for keyword extraction. Finally, as described in section 3, sentiment analysis is conducted. Then web scraping is executed in section 4 to discover appropriate subreddits in this example to establish the target audience for the input video, combining the results from all three phases.

1) Video Analysis: When a user inputs a video, a video feature label extraction method is employed. This method is used to extract feature labels from a video automatically. The basic characteristics are produced by gathering video quality and then capturing key frames of video. This approach consists of a two-stream ConvNet and a revolutionary automated an- notation architecture capable of consistently annotating crucial frames in a movie for the ConvNet's self-supervised learning. To recognise unique frames, the suggested ConvNet learns deep appearance and motion properties. The trained network can then recognise crucial frames in videos.

By merging the visual and motion feature vectors out of each sequence frame to form a composite representation, the representation capacity is improved. It makes use of Linear Discriminant Analysis (LDA) [15] and [16] to evaluate how distinct a certain image in a sequence of pictures in a video is (LDA). It creates a matrix VA by fusing the attributes of all videos belonging to the same class A, as well as V1, V2,... and additional features from videos in other categories. LDA is also employed in the learning of C projected vectors. The LDA projection matrix reduces Class A within-class distance while increasing Class B between-class distance (class B).

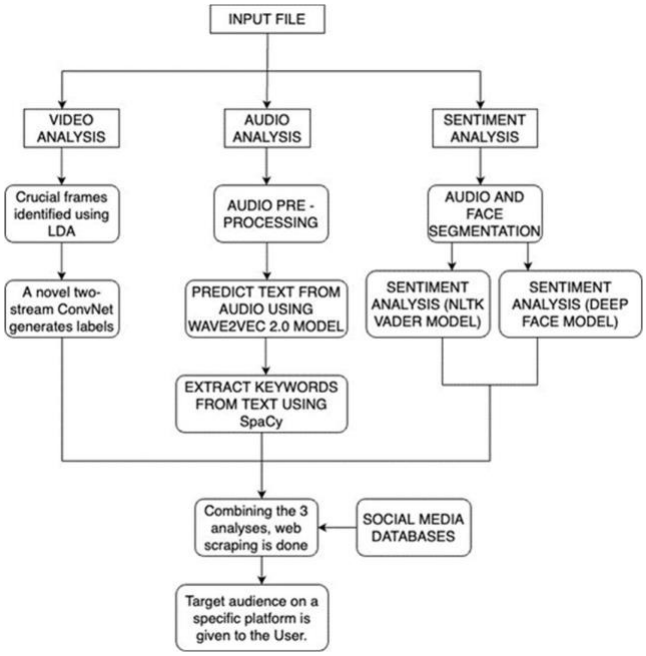

Figure 1: Proposed Methodology

$$VA = V1 \qquad (1)$$

$$VB = [V2, V3, ..., VC] \qquad (2)$$

LDA can be employed to generate the projecting vectors WA. (Given VA and VB)

$$WA = LDA(VA, VB) \quad (3)$$

For each training video, a frame score (label value) is computed. A low score correlates to a minor alteration between frames, implying that the activity in that frame is insignificant. Now, each frame score (label value) for each class A training video is computed using WA from Equation 1, Equation 2 and Equation 3:

$$f_{i,m} = ||F_{i,m} - WAWT\ F_{i,m}||2 \qquad (4)$$

In Equation 4, Fi, m is the feature vector of the i-th frame of class A's m-th video, and fi, m is just a value indicating how distinct it is. The preceding is a two-class LDA technique which is executed for every category. For instance, if C = 20, this will continue the operation for every one of the 20 classes, calculating the feature value of every frame within every training video for every class employing Equation 4

Learning a deep model to replicate the very same result gets straightforward after creating a label fi, m for every frame within every training video presented previously. Due to the fact

that fi, m is a continuous variable instead of a discrete variable, the proposed approach requires train a regression model as opposed to a classification structure. The training videos' optical flow and appearance data are once more assessed using a two-stream CNN, and the resulting features are combined exactly as previously. The suggested framework is a deep two-stream convolutional neural network, with S1 expressing the appearance stream and S2 indicating the optical flow stream.

The network architecture for these two streams is the same as that of the VGG-16. The significant distinction is that the subsequent layers (fc7, fc8, and softmax) are eliminated sequentially, starting with the second completely connected layer (fc6). The completely linked layer fc6 from the appearance and motion streams is labelled as fc6-1 and fc6-2 for clarity. The results of the fc6-1 and fc6-2 layers, which were followed by the new fc7 and fc8 wholly linked levels, are combined to create the input of the new completely connected layer of this deep two-stream ConvNet. The ground truth is used to calculate the Euclidean loss in the final layer, which is a regression layer.

To identify key frames, the completely trained model explained above is fed an input video V, and the prediction outputs are indicated as [S1, S2,...SM]. To fit a curve to the outputs [S1, S2,...SM], a smooth spline function is utilised. The frame score or label fi, m indicates to how equivalent a frame is to other frames in its category, as indicated by Equation 4.

This collects motion and appearance attributes from raw video frames as well as the optical flow images linked with them. The frame and optical flow picture fc7-layer charac- teristics are retrieved and synthesized. These are utilised to generate frame level score values and, as a result, dynamically label/annotate frames using LDA. Labels are made in this manner.

2) Audio Analysis: Machine intelligence and deep learning are increasingly being used in audio analysis. To train a statistical or ML model, significant properties from an audio stream are first retrieved. Audio feature extraction is a critical stage in audio processing, that is a subset of signal processing. It is concerned with the manipulation or modification of audio signals. By translating analog and digital signals, it reduces undesirable distortion and adjusts the time-frequency spectrum. It is focused with computational sound manipulation techniques.

The Wav2Vec2.0 model is a new addition to the Hugging- Face transformers' most recent iteration and has the ability to resolve audio-related Natural Language Processing (NLP) problems. Text is retrieved from audio in this paper and then the text's keywords are extracted as the NLP challenge.

2.1) Audio Preprocessing: The size of the videos is con- ventionally huge, and processing these audio files is a lengthy process. As a result, the audio file is divided into smaller chunks to improve the efficiency of the suggested approach. For example, if the video is 15 minutes long, it is divided into three portions of 5 minutes each. As a result, these parts can be processed simultaneously.

2.2) Predict Text from Audio: The method uses the Wav2vec 2.0 model to predict text from audio. Wav2Vec 2.0 is one of the most modern state-of-the-art models for Automated Voice Recognition thanks to a self-supervised training technique that is relatively novel in this field. Twice the model is trained. The best possible speech representation is sought after in the first phase, which is carried out in self- supervised mode with unlabeled input.

In the second stage of training, supervised fine-tuning, the model is trained to predict particular words or phonemes using labelled data. In Wav2Vec 2.0, the first method of transformation is the transformer layers, and the second approach is quantization. The goal is to explore such a context representation (ct) for a latent masking representation (zt), in more detail. Because this is a classification problem, it appears that a softmax function is the natural choice for selecting the most suitable keywords. In this approach, a Gumbel softmax shown in Equation 5 has been used.

$$P_{g,v} = \frac{\exp\left(sim(l_{g,v}+n_v)/T\right)}{\sum_{k=1}^{V}\exp\left(sim(l_{g,v}+n_v)/T\right)} \qquad (5)$$

where,

- sim = cosine similarity,
- l ∈RGxV = logits calculated from z,
- nk = -log(-log(uk)),
- uk = sampled from the uniform distribution U(0, 1),
- T = temperature

Randomization and temperature are now two added features. As a result of randomization, the model is more likely to select alternate code words and then vary their weights during training. It is critical, especially at the start of training. The effect of randomization diminishes over time as the tempera- ture lowers from 2 to 0.5. It preprocesses the raw waveform with convolutional layers before applying a transformer to augment the voice representation with context. The result may be obtained by giving the smaller audio files to the model and looping them. Because this model was trained on a frequency of 16KHz, the input audio should likewise be of that frequency.

2.3) Extract keywords from text: After evaluation and consideration of multiple models namely Rake, Bert, Yake, and SpaCy; the model that complements the suggested methodology the best is SpaCy. This can be concluded from Table I. It can be seen that SpaCy is 20 times and 443 times faster than NLTK in the case of tokenization and tagging. This newer software than NLTK or Scikit-Learn aims to make deep learning for text data analysis as simple as feasible. The processes for extracting keywords from a text using spacy are as follows.

- Tokenize the inputted text information.
- The trending words should be extracted from the token list.
- Set the popular words to the pos tags. PROPN (proper noun), ADJ (adjective) and NOUN (noun) (The POS tag list can be customized.)
- Find the T most common frequent and trending words from the list.
- Print the outcomes

When using spaCy, passing a text string to an NLP object is the first thing you should do. The input text string must pass through this object, which is essentially a pipeline of several text pre-processing procedures.

Figure 2: SpaCy pipeline

It can be seen in Figure 2 that the NLP pipeline has multiple components such as parsers, taggers, Named Entity Recognition etc. So the text firts passes through the pipeline before it can be used.

3) Sentiment Analysis: In this approach, the first stage in the analysis will be to segment the input video clip based on the amount of quiet or frequency of the conversation. The audio and face sentiment analysis will be performed individually on these portions, and the results will be integrated that may be utilised for further research.

To get segmentation points based on quiet and background noise, the input video clip is first converted to a .wav file using the ffmpeg package. Followed by the conversion, the segmentation is carried out with the help of the ”my-voice- analysis” library’s myspsr() function, which takes an audio file as input and generates a “.Textgrid” file for the video, with the time range of the audio file with constantly sounding or quiet period. My-Voice-Analysis is a Python module for voice analysis (simultaneous speech, high entropy) that does not require transcription. It recognises syllable boundaries and breaks speech. This textGrid file is then used to separate the video into parts for face and audio analysis.

The video, like the audio, is separated into many pieces with the same time range depending on segmentation. Multiple frames are retrieved from these chunks and tested to see if a face shows in the selected frame. In OpenCv, the ”haarcascade classifier” is used for validation. Deepface, a Python frame- work for face recognition and facial attribute analysis, is then used to evaluate the sentiment of face images. It’s a hybrid facial recognition framework that incorporates cutting-edge models like VGG-Face, Google Facenet, OpenFace, Facebook Deepface, ArcFace, and Dlib. These models have already acquired and surpassed human-level precision. The package is mostly based on Keras and TensorFlow. Deepface examines facial characteristics such as facial expression (including 7 different emotions). This function returns the emotion iden- tified in the picture together with the likelihood score which is shown in Figure 6.

4) Web Scrapping: Following the combination of key- words, labels, and sentiment analysis, web scraping is per- formed. It is accomplished by the use of several APIs, such as Tweepy for Twitter and Praw for Reddit. Because Reddit has various communities, the web scraping algorithm can assist to scrape select groups where viewers may be interested in viewing the video being analysed. In the case of Twitter, the scrapper aids in the discovery of relevant tweets, which boosts user engagement if these tweets are supplied with the video, hence enhancing retention. These stages aid in strengthening retention and growing the product by improving marketing and commercial strategies.

Table I: Comparison of spaCy with NLTK

	Absolute (ms per doc)			Relative (To spaCy)		
System	Tokenise	Tag	Parse	Tokenise	Tag	Parse
spaCy	0.2ms	1ms	19ms	1x	1x	1x
NLTK	4ms	443ms	n/a	20x	443x	n/a

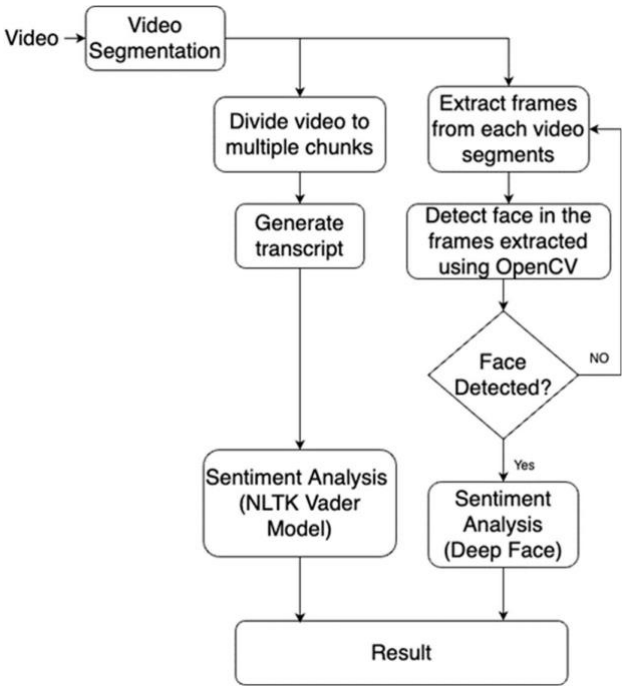

Figure 3: Sentiment Analysis

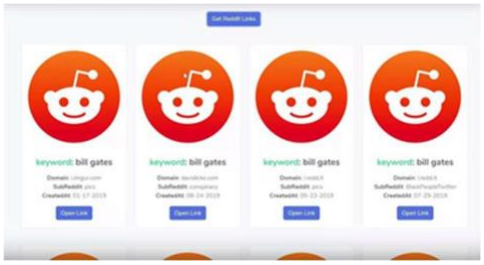

Figure 4: Web scraping

Table II: Results on Video analysis

Input File	Labels	Keywords
This is a video of a speech given by Bill Gates during the inauguration of a Microsoft event.	person, man, indoor, people, kitchen, tree, group, player,monitor, standing, outdoor, glasses, headdress, laptop, suit, wall, necktie, computer, booth, sign, display, television,posing, wearing, sky, screen, electronics, floor, eating, smiling, clothing	bill gates, bill, order,microsoft, estimated net, world, money,gates, conclusions

V. RESULTS

The paper presents a model for business development and marketing by providing a targeted audience through web scrapping. To demonstrate some results, the model was run on the following videos presented in Figure 5. The suggested ConvNet, uses motion information and deep appearance from video frames to distinguish and identify frames. Using LDA, it is determined how unique the frames from the video are. The frame score is then determined as explained in section 3. The main goal is to have the keywords, labels and the sentiment analysis generated by the algorithm and output the best possible target audience. The algorithm's output is shown in Figure 4, Figure 6 and Table II. The model has developed the intended Reddit groups to be focused by combining keywords, labels, and sentiment analysis, followed by web scraping.

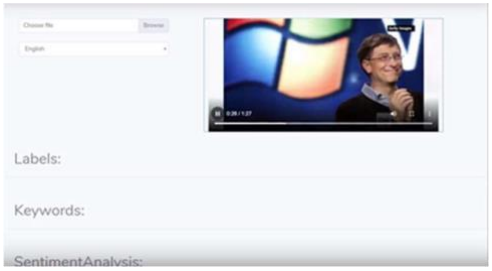

Figure 5: Input Video

Figure 6: Labels and Keywords Generation

V. CONCLUSION AND FUTURE SCOPE

The potential of digital enterprises can only be unlocked by targeted digital marketing. Any local or internet business can develop more quickly and effectively this way. This proposed method of combining video analysis, audio analysis and sentiment analysis gives businesses the chance to quickly decide on important matters using vast data. The generation of labels using a two-stream ConvNet that employs the LDA concept, keywords using Wav2vec2.0 model, and followed by web scraping is critical because it improves data analysis accuracy, allows you to analyse more data in less time, adjusts to new data and changes, and allows you to automate marketing operations and eliminate monotonous work. The model that is presented in the study seeks to achieve these goals. There are a lot of possibilities for future scope in this project. Some options include browser extensions, such as a Chrome extension that enables you to navigate to the section of a video that has the information you want by simply inputting a keyword or a YouTube add- on. The concept provided might not only be applicable to YouTube video producers but also to a wide range of other entities, including businesses getting ready to launch new goods and individuals seeking to make a point in an interview or on social media. The model could also potentially accommodate a number of native tongues. The possibility of incorporating Facebook Graph into the app by circumventing privacy restrictions is exciting. Thus, the paper presents a one-stop shop for video analytics.

VI. ETHICAL STATEMENT

We authors, declare that all sources have been dully cited in the manuscript.

VI. AUTHOR'S CONTRIBUTION

In this paper, six authors have contributed. Dhruv Mehta developed the video analysis algorithm and worked on the formatting of the final draft. Harvy Gandhi documented the literature review and executed the technical tasks. Jinay Parekh assisted in coding of the algorithm and finalized the method- ology. Tanya Sharma worked on the overall documentation of the paper and tested various machine-learning models. Pratik Kanani and Aniket Kore worked on overall supervision.

VII. CONFLICT OF INTEREST

Authors do not have any conflict of interest.

REFERENCES

[1] A. Miklosik, M. Kuchta, N. Evans, and S. Zak, "Towards the adoption of machine learning-based analytical tools in digital marketing," Ieee Access, vol. 7, pp. 85 705– 85 718, 2019.

[2] A. G. Tettamanzi, M. Carlesi, L. Pannese, and M. San- talmasi, "Business intelligence for strategic marketing: Predictive modelling of customer behaviour using fuzzy logic and evolutionary algorithms," in Workshops on Applications of Evolutionary Computation. Springer, 2007, pp. 233–240.

[3] J. Salminen, V. Yoganathan, J. Corporan, B. J. Jansen, and S.-G. Jung, "Machine learning approach to auto- tagging online content for content marketing efficiency: A comparative analysis between methods and content type," Journal of Business Research, vol. 101, pp. 203– 217, 2019.

[4] V. A. Brei et al., "Machine learning in marketing: Overview, learning strategies, applications, and future developments," Foundations and Trends® in Marketing, vol. 14, no. 3, pp. 173–236, 2020.

[5] E. W. Ngai and Y. Wu, "Machine learning in marketing: A literature review, conceptual framework, and research agenda," Journal of Business Research, vol. 145, pp. 35– 48, 2022.

[6] D. Bumblauskas, D. Gemmill, A. Igou, and J. Anzen- gruber, "Smart maintenance decision support systems (smdss) based on corporate big data analytics," Expert systems with applications, vol. 90, pp. 303–317, 2017.

[7] R. Raturi, "Machine learning implementation for busi- ness development in real time sector," International Jour- nal of Pure and Applied Mathematics, vol. 119, no. 15, pp. 1289–1300, 2018.

[8] M. Zorrilla, J. Flo´rez, A. Lafuente, A. Martin, J. Mon- talba´n, I. G. Olaizola, and I. Tamayo, "Saw: Video analysis in social media with web-based mobile grid computing," IEEE Transactions on Mobile Computing, vol. 17, no. 6, pp. 1442–1455, 2018.

[9] A. Janchevski and S. Gievska, "A study of different models for subreddit recommendation based on user- community interaction," in International Conference on ICT Innovations. Springer, 2019, pp. 96–108.

[10] A. Aggarwal, B. Gola, and T. Sankla, "Data mining and analysis of reddit user data," in Cybernetics, Cognition and Machine Learning Applications. Springer, 2021, pp. 211–219.

[11] D. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," in Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 4. IEEE, 2000, pp. 167–170.

[12] J. Yuen, B. Russell, C. Liu, and A. Torralba, "Labelme video: Building a video database with human annota- tions," in 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009, pp. 1451–1458.

[13] S. Kulhare, S. Sah, S. Pillai, and R. Ptucha, “Key frame extraction for salient activity recognition,” in 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016, pp. 835–840.
[14] G. Guan, Z. Wang, S. Lu, J. Da Deng, and D. D. Feng, “Keypoint-based keyframe selection,” IEEE Transactions on circuits and systems for video technology, vol. 23, no. 4, pp. 729–734, 2012.
[15] J. Ye, R. Janardan, and Q. Li, “Two-dimensional linear discriminant analysis,” in Advances in Neural Informa- tion Processing Systems, L. Saul, Y. Weiss, and L. Bot- tou, Eds., vol. 17. MIT Press, 2004.
[16] S. J. Prince and J. H. Elder, “Probabilistic linear dis- criminant analysis for inferences about identity,” in 2007 IEEE 11th international conference on computer vision. IEEE, 2007, pp. 1–8.