European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 04 , 2020

# BEHAVIOUR ANALYSIS OF SOCIAL NETWORK APPLICATION USING NATURAL LANGUAGE PROCESSING – A MACHINE LEARNING APPROACH

**S. Godfrey Winster[1], A. Siva Kumar[2], P.V.Gopirajan[3], V.Loganathan[3]**

[1]Department of Computer Science and Engineering, SRM Institute of Science & Technology, SRM Nagar, Kattankulathur- 603203, Chengalpattu District, Tamil Nadu, India.

[2]Department of Information Technology, Sri Sairam Engineering College, Sai Leo Nagar, West Tambaram,Chennai-600 044, Tamil Nadu, India.

[3]Department of Computer Science and Engineering, Saveetha Engineering College, Saveetha Nagar, Thandalam-602 105, Tamil Nadu, India.

**Abstract:**

**World Wide Web (WWW) has become a monstrous wellspring of client delivered content and opinionative data. Utilizing web-based social networking, for example, Twitter, Facebook, LinkedIn and so forth client share their points of view, feelings in a helpful way, where a great many individuals express their perspectives in their everyday connection, which can be their conclusions and assessments about specific thing. These routinely creating unique data are, beyond question, a staggeringly rich wellspring of information for such an essential administration process. To mechanize the examination of such information, the region of Sentiment Analysis (SA) has developed. It targets recognizing opinionative data in the Web and gathering them as demonstrated by their furthest point. This is achieved by introspecting the various data uploaded by the user. These data are categorized under different emotions such as social awareness, curiosity, emotions, creative, advertising and so forth., with the assistance of the keywords used in the data uploaded. In that conduct dissect we are utilizing content based separating, Collaborative Filtering and Natural Language pre-processing calculation. Relies on this calculation we will order the users.**

## 1. Introduction:

The 21st century represents a period of development in many fields of study (e.g., therapeutic science, software engineering, monetary, Image processing, human science, and so forth.) [1]. Huge advances have been accomplished and made effects our living condition. One new marvel Sharing Economy that rethinks the proprietorship among clients gets wide discourses. Conveyance administration, for example, bundle, mail, and nourishment, may best depict the

situation [2]. A significant point under this situation is decentralization. The specialist co-op relocates from a firm to an individual, and another thing to a current (only here and there utilized) thing. The above case just demonstrates a straightforward case, and it very well may be generally applied to all the lingering assets, regardless of as substantial or immaterial, around our living condition with the help of urban innovation [3]. Human practices can be watched through, and under, the sharing economy. Issues, for example, trust, inclination, feeling and basic leadership concerning human conduct have pulled in huge considerations from specialists as of late. Various sorts of datasets become accessible and explicit behaviour(s) may likewise be checked by information gathered from sharing economy stages (e.g., convenience sharing, vehicle sharing, nourishment sharing, egotistical sharing, and so forth.) [4–6].

Taking two reads above for example researchers have called attention to that the outward appearances may cause impact on the distinction on choice on sharing economy stage that requires high communication with clients. Various articulations (e.g., delight, furious, pitiful, and unbiased) on close to home profile picture may prompt various emotions just as buy expectation to potential clients significantly other target factors, for example, star-rating and remarks, are sure/negative. The other case structured a system for execution of client based movement as indicated by station-based single direction vehicle sharing [7,8].

Clients, therefore, have indicated some certain practices while utilizing such sharing economy stages and these conduct information might be another channel to accomplish better comprehension of end clients. Mining [9,10], investigation and human practices is an approach to accomplish client understanding. The better we can comprehend the objective clients, the more we can serve them. Human practices can chiefly be isolated into outside or inward settings. A basic definition is that outer settings are factors from outside, for example, culture, gadget, internet based life, and time while interior settings are those elements from inside, for example, nationality, sex, age, inclinations or even related involvements. Every one of these settings may prompt the progressions on basic leadership process [11,12]. For example, clients may have extra worries because of their age and sexual orientation, or clients may change their ready to lease explicit sort of settlement or vehicle as per the circumstance. Computer vision based systems [13] were still on demand and being used in various sectors including image identification of data, optical character recognition and image detection in moving objects [14].

This marvel certainly demonstrates that occasionally clients may settle on choices deliberately or inadvertently dependent on interior/outer settings, and accordingly mining such settings to have better comprehension may turn into an earnest issue particularly in the situation of sharing economy. Clients may have comparative practices on the off chance that they share comparative settings, regardless of from inside or outer, in like manner. Access to these normal practices is valuable for specialists, just as organizations, to give better administrations versatile to explicit clients than without it. For example, uses' practices might be diverse as per their nation, sexual orientation, age, and so on. Consequently, so as to find and recognize those understood human practices, Artificial Intelligence techniques like Decision Tree [15], Support Vector Machine [2,16], Neural Network [17,18], Apriori and so forth., were regularly applied to meet specific purposes. Yet, notwithstanding, the greater part of the learning results are relied upon to get shrouded designs that are beforehand uninformed [19,20].

Subsequently, better administrations are forthcoming if human practices can be demonstrated. Considering the developing prominence in sharing economy this examination endeavours to break down basic leadership process through a widespread model that supports dynamic and group detecting and investigation of client produced content just as settings

adequately. The conduct information extricated from the client produced substance and its related settings is particularly thought. This investigation especially takes a gander at the inward factors (e.g., nationality, sex, and age) and outer components (e.g., gadget, internet based life, and time) that ponder certainly the distinction group's inclination and conduct.

The field of AI has consistently imagined machines having the option to emulate the working and capacities of the human psyche. Language is considered as one of the hugest accomplishments of people that has quickened the advancement of mankind. Thus, it's anything but an unexpected that there is a lot of work being done to incorporate language into the field of AI as Natural Language Processing (NLP) [21–23]. Common techniques used in NLP includes, Named Entity Identification, Aspect Mining, SA, Script Summarization and Subject Modelling [24].

This paper plots the way toward demonstrating at dynamic level, and afterward, concocts a solid model for confirmation the practicality. Various sorts of datasets become accessible and explicit behaviour(s) may likewise be checked by information gathered from sharing economy stages. At that point a lot of AI calculations are executed for recover the highlights on clients' inclination just as conduct. We use WEKA Tool [25,26] for precision and proficiency which gives the weightage for the information. Following the presentation, related thinks about are condensed in a definition and structure on the widespread model. The discoveries from the information Section VI at that point finishes up the work.

## 2. Related work:

In [27], they utilized a supposition mining estimation investigation calculation. This approach urges us to arrange the inclination into different thought. The focal points are augmentation in the sufficiency, it is precise in nature, clear and reproducible, and incorporates recognizing, dissecting, consolidating, thinking, and itemizing the evidence. What's more, the restrictions are Poor information: Efficiency of the revelation procedure and the nature of the found information are emphatically reliant on the nature of the information. The intricacy of the strategy has likewise expanded.

Investigational outcomes that applied Support Vector Machine (SVM) discussed in [28] describes about a benchmark database to prepare an opinion classifier. The focal points are N-grams and diverse weighting plan, for example, chi-square were utilized to separate the most old style highlights. It is characterized by an arched advancement issue for which there are a proficient technique. The confinements of SVM is predominantly identified with the choice of part. Just two distinct classifiers have been executed and looked at.

In [29], they portrayed the endeavour to tackle the issue of handling two classifier by making the greatest utilization of both the old-area information. To get data from the new territory data, they proposed Adapted Naive Bayes. This has the focal points that it could improve the introduction of base classifier definitely, and even give a lot of best execution over the trade learning design. The limitations of the Adapted Naive Bayes are that there will be no occasion of class mark and a particular attribute regard together. The repeat based probability estimation will be zero. In course of action of undertaking you need a significant dataset in order to make a strong estimation of the probability of each class.

Open source software was used by [29] depicts the expounding the various methodologies of SA and Opinion Mining for numerous dataset and discovers which approach is best for which dataset. They have used R apparatuses of various occasions from twitter and did pre-preparing and compute slant scores from that occasions. The preferences are this backings

augmentations and performs wide assortment of capacities, for example, information control, factual demonstrating and designs. Designers can without much of a stretch compose their own product and disperse it as extra bundles. The impediments are that the client must settle on the investigation grouping and execute it bit by bit. The default Windows and Mac OS X graphical UI is constrained to basic framework collaborations and does exclude factual strategies.

Machine learning approach was implemented in [30] to detect the spread of fake news in twitter social media. They have converted the tweets published in Chinese language to English language and formed those collections of words as a dataset. Further, they have applied NLP algorithm to segregate the offensive words from the tweets. They have achieved a good accuracy in word prediction using NLP.

Tweets against racism was addressed by [31] in U.S. They have collected tweets related to racism and segregated the words related to racism and formed a huge master dataset. They have used SVM approach of machine learning and carried out sentiment analysis to predict the words related to race. They have suggested the possibility of removing negative tweets after achieving high accuracy over 90 percent.

A similar approach to [31] was carried out in [32] for predicting the Arabic offensive language on the popular Twitter network. In addition to [31], here they have used lexical analysis for word segregation and formed them as a dataset. They have also achieved good precision with this approach of machine learning.

Sentiment analysis was performed in [33] for predicting the popularity of the mobile application present in the play store. Authors collected data from Facebook as a form of graph and analyzed the data with sentimental words present in users post and predicted the accuracy of the mobile application.

## 3. Experimental analysis:

In proposed framework 3 algorithms were utilized. Collaborative filtering, Content Based Filtering and NLP. Relies on the client posts and exercises it will order and examinations the client practices. Trust, inclination, feeling and basic leadership concerning human conduct have pulled in noteworthy considerations from analysts as of late. Various sorts of datasets become accessible and explicit behaviour(s) may likewise be checked by information gathered from sharing economy stages. Content based separating calculation used to dissect the Social Awareness, Innovative, Marketing practices. The coding language utilized here is java since it is straightforward, versatile, dynamic, and secure and has elite. The database utilized is MYSQL. The points of interest are Collaborative Filtering motors work best when the client space is huge (since that is their wellspring of information). Content Based Filtering motors will in general give essentially things more inside the "cleared street" of client tastes, however there are counter measures to expand assorted variety. NLP is utilized to investigate syntactic/correspondence for post. Community oriented separating calculation used to break down the Emotional Quotient and Foul Languages. We additionally use WEKA Tool for precision and productivity which gives the weightage for the information.

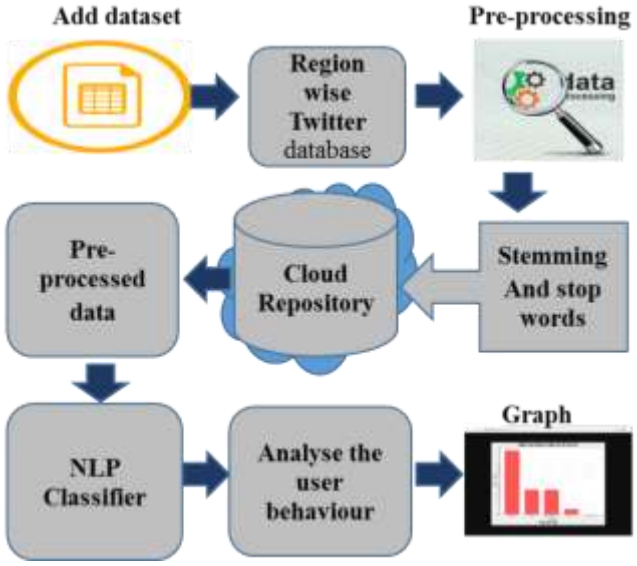


**Fig.1. Overall schematic diagram of this study**

Fig.1 demonstrations that how the behaviour analysis has been done in this study. The dataset has been gathered from the tweeter database from the open source and afterward that database is put away in mysql. The following stage is pre-preparing the gathered information. The pre-preparing has two procedure for example expulsion of stemming and stop words. At that point the grouping will be done on the tweets utilizing the watchwords and the last yield will result as a graph.

## 4. Dataset and module description:

In this section IV we depict about the Implementation, data acquired and the example code for thefeeling examination. The significant modules are portrayed in this area.

### 4.1 Data collection:

In this movement, information noteworthy to the examination task are recuperated from the database. Information assurance is described as the route toward choosing the correct information type and source, similarly as sensible instruments to accumulate information. Information assurance goes before the genuine act of data collection. In the first module, any of the administrator need to login, they ought to login by giving their name and secret key. Legitimate username and password implies divert to administrator landing page. Administrator transfer the .xls dataset. After transfer the dataset xl data are brought into Mysql database. In this dataset contain clients' tweets.

### 4.2 Data pre-processing :

Machine Learning warehouse is an assortment of databases, area speculations, and information generators that are utilized by the AI accessible group for the exact investigation of AI calculations

**Table 1: NLP Keyword identifiers**

| Identifier | **Description** |
|---|---|
| Name | (person name) |
| Gender | (Male or Female) |
| Tweets | (Opinion of a person) |
| Location | (location of person who tweets) |
| Sentiment | (The accuracy rate in number) |
| Role | (Role of a person) |

Table 1 shows the sample identifiers used in this study for NLP

### 4.2.1 Stop Words:

The target file content is tokenized and individual words are taken care of in group. A single forestall word is scrutinized from stop word list. The stop word is appeared differently in relation to target message in sort of display using back to back hunt system. In case it organizes, the word in show is cleared, and the assessment is continued till length of group.
After clearing of stop word absolutely, another forestall word is examined from stop word list and again computation runs reliably until all the Stop words were broke down.

### 4.2.2 Stemming:

Stemming is process which allows the system to predict the wrong/misspelled/misleading and meaningless words and substitute the meaningful word from the available spelling. Words with identical meaning were grouped as a root and formed as stem. Stem formed with the help of NLP were formed as a base or root structure. Dataset was created for stems with various word combinations with the help of dictionary. Dataset created was given to the database for runtime comparison and retrieval of identical and meaningful words.

### 4.3 Behaviour analysis:

There are two kinds of direct examination that can be used for evacuating models depicting huge classes or to envision future information designs. These two structures are according to the accompanying

- •Classification
- •Prediction

### 4.3.1 Classification:

Classification predict full scale class stamps; and estimate models foresee constant regarded limits. For example, we can manufacture a request model to orchestrate bank acknowledge applications as either secured or hazardous, or an expectation model to predict the utilizations in dollars of likely customers on PC gear given their pay and occupation.

### 4.3.2 Prediction:

In this module we are using 3 algorithms to analyse user behaviour.

- Content based Filtering

- Collaborative Filtering
- Natural Language Processing (POS Tagger)

#### 4.3.2.1 Content based Filtering:

A recommender system, that is, a proposed system (some of the time supplanting "system" with a corresponding word, for example, platform and engine) is a subclass of data scrutinising framework that looks to anticipate the rating and inclination that a client would provide for an input information condition. In this module we foresee the client conduct for Social mindfulness, Innovative thoughts and business related posts. Content based separating algorithm characterize the user present related on inventive, legislative issues and business.

#### 4.3.2.2 Collaborative Filtering:

In the more broad sense, collaborative filtering is the way toward scrutinizing for data and the samples utilizing methods comprising coordinated effort among various authorities, perceptions, data sources, and so on. Most cooperative separating frameworks apply the supposed neighbourhood-based procedure. In the area based methodology various clients is chosen dependent on their closeness to the dynamic client. An expectation for the enthusiastic remainder and foul dialects we are utilizing community oriented sifting calculation. It is made by figuring a weighted normal of the post of the chose users.

#### 4.3.2.3 Natural Language Processing (POS Tagger)

In NLP, grammatical form labelling (POS labelling or POS labelling or POST) were termed as syntactic labelling or word-class disambiguation. It delivers the possibility to escalate the total number of word present in the book (corpus) as relating to a specific grammatical feature present to be searched, among these two of its definition and its unique situation. That is, an NLP helps in accessing the associated words with contiguous and related words present with its own expression, sentence of particular text or a given paragraph. This NLP methodology as practiced with school kids in school education can be taken as the sample to choose things, action words, descriptors, verb modifiers, and so forth. This NLP calculation procedure is applied in this study and utilized its effectiveness to get the keywords from the social networking site chosen in this study.

## 5. Output:

These modules listed here were implemented as a prototype with the general implementation strategy and analysed their efficiency. Fig.2. shows the data upload module’s screenshot. This shows how the data is uploaded for behaviour analysis.


**Fig.2. Data Upload module**

**Fig.3. depicts the screenshot for the removal of the stop words.**

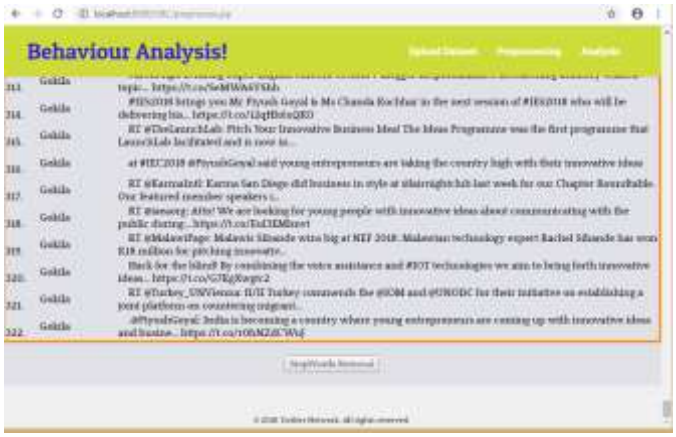


**Fig.3. Stop Words**

**Fig.4. shows the screenshot for the removal of stem words. This shows the removal of stem words and leave the root words for the given database.**


**Fig.4 Stemming**

Fig.5 shows the screenshot for the processing of the behaviour analysis. This field allows the user to select one particular person to get the result of behaviour analysis.


**Fig.5 Behaviour Analysis Process**

Fig. 6 shows the screenshot for the result for the behaviour analysis. This gives the result of the selected user's behaviour based on his/her tweets. It gives different emotions of the person using the keywords.


**Fig.6. Behaviour Analysis Result**

Fig.7. portrays the screenshot for the chart generation. This generate the chart for the result that has been analysed.

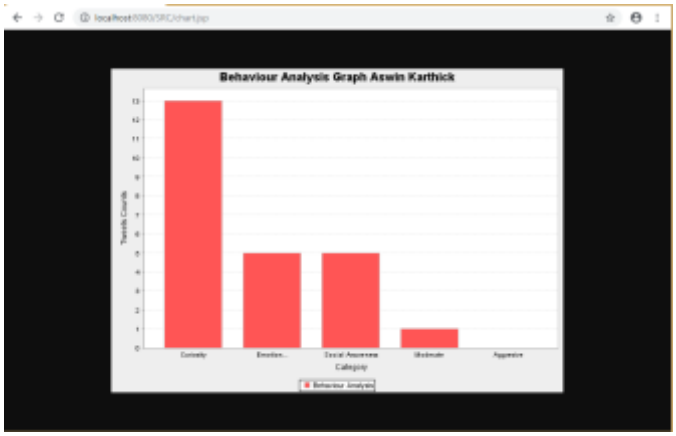


**Fig.7. Chart Generation**

Fig.8 shows the screenshot for the result of the WEKA tool. This tool gives the weightage for the data present and the accuracy and efficiency.

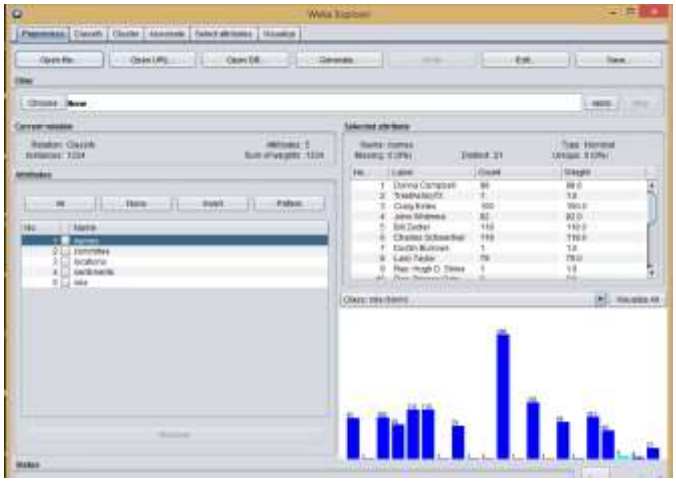

**Fig.8. WEKA Tool**

## 6. Conclusion:

Accomplishment of sharing economy makes numerous open doors for the scientists in all Fields. From the perspective of software engineering, the stages that actualize sharing economy produced a colossal number of information, particularly client created information that can be utilized to accomplish better comprehension of people. With this worry, this paper plans an all-inclusive client model that supports the dynamic group detecting on client inclination, and investigation of associated basic leadership conduct. This examination particularly focuses on the examination of connections between inward factors (e.g., nationality, sexual orientation, and age) and outer variables (e.g., gadget, online networking, and time). Results signify that maximum of the people's tweets are constructed on inquisitiveness, social cognizance and emotions.

## 7. References:

[1] Shanmugaprakash M, Sivakumar V. Development of experimental design approach and

ANN-based models for determination of Cr(VI) ions uptake rate from aqueous solution onto the solid biodiesel waste residue. Bioresour Technol 2013;148:550–9. https://doi.org/10.1016/j.biortech.2013.08.149.

[2] Lieder M, Asif FMA, Rashid A. A choice behavior experiment with circular business models using machine learning and simulation modeling. J Clean Prod 2020;258:120894. https://doi.org/10.1016/j.jclepro.2020.120894.

[3] Lei K, Du M, Huang J, Jin T. Groupchain: Towards a Scalable Public Blockchain in Fog Computing of IoT Services Computing. IEEE Trans Serv Comput 2020;13:252–62. https://doi.org/10.1109/TSC.2019.2949801.

[4] Sithole MPP, Nwulu NI, Dogo EM. Dataset for a wireless sensor network based drinking-water quality monitoring and notification system. Data Br 2019;27:104813. https://doi.org/10.1016/j.dib.2019.104813.

[5] Hammoumi A, Moreaud M, Ducottet C, Desroziers S, Hammoumi A, Moreaud M, et al. Distance transform data augmentation and stochastic patch-wise image prediction methodology for small dataset learning To cite this version : HAL Id : hal-02879709 2020.

[6] Lumini A, Nanni L. Fair comparison of skin detection approaches on publicly available datasets. Expert Syst Appl 2020;160:113677. https://doi.org/10.1016/j.eswa.2020.113677.

[7] PU J, LI G, CAO L, WU Y, XU L. Investigation and analysis of the psychological status of the clinical nurses in a class A hospital facing the novel coronavirus pneumonia. Chongqing Med 2020;49:E015–E015.

[8] Fessell D, Cherniss C. COVID-19 & Beyond: Micro-practices for Burnout Prevention and Emotional Wellness. J Am Coll Radiol 2020. https://doi.org/10.1016/j.jacr.2020.03.013.

[9] Kaspa LP, Akella VNSS, Chen Z, Shi Y. Towards Extended Data Mining: An Examination of Technical Aspects. Procedia Comput Sci 2018;139:49–55. https://doi.org/10.1016/j.procs.2018.10.216.

[10] Duong LVT, Thuy NTT, Khai LD. A fast approach for bitcoin blockchain cryptocurrency mining system. Integration 2020;74:107–14. https://doi.org/10.1016/j.vlsi.2020.05.003.

[11] Corrales DC, Ledezma A, Corrales JC. A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks. Appl Soft Comput J 2020;90:106180. https://doi.org/10.1016/j.asoc.2020.106180.

[12] Su CW, Qin M, Tao R, Umar M. Financial implications of fourth industrial revolution: Can bitcoin improve prospects of energy investment? Technol Forecast Soc Change 2020;158:120178. https://doi.org/10.1016/j.techfore.2020.120178.

[13] Karthikeyan V, K GP, Siva R, Gopirajan P V. Computer vision based Comparative Studies on the Physicochemical Analysis and Bacterial Biota in Different Milk Samples. Int J Res Pharm Sci 2020;11:3699–703.

[14] Padmapriya K, Gopirajan P V, Kumar KS. Occlusion Detection with Background Elimination and Moving Object Tracking 2019:9248–52. https://doi.org/10.35940/ijrte.D9287.118419.

[15] R PT. A Comparative Study on Decision Tree and Random Forest Using R Tool. Ijarcce 2015;4:196–9. https://doi.org/10.17148/ijarcce.2015.4142.

[16] Chen H, Xu L, Ai W, Lin B, Feng Q, Cai K. Science of the Total Environment Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy 2020;714. https://doi.org/10.1016/j.scitotenv.2020.136765.

[17] Gitifar V, Eslamloueyan R, Sarshar M. Experimental study and neural network modeling of sugarcane bagasse pretreatment with H2SO4 and O3 for cellulosic material conversion to sugar. Bioresour Technol 2013;148:47–52. https://doi.org/10.1016/j.biortech.2013.08.060.

[18] Wang X, Girshick R, Gupta A, He K. Non-local Neural Networks. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018. https://doi.org/10.1109/CVPR.2018.00813.

[19] Antonopoulos I, Robu V, Couraud B, Kirli D, Norbu S, Kiprakis A, et al. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. Renew Sustain Energy Rev 2020;130:109899. https://doi.org/10.1016/j.rser.2020.109899.

[20] Santosh KC. AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. J Med Syst 2020;44:93. https://doi.org/10.1007/s10916-020-01562-1.

[21] Woodward K, Kanjo E, Oikonomou A, Chamberlain A. LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. Pers Ubiquitous Comput 2020. https://doi.org/10.1007/s00779-020-01427-x.

[22] Vishwakarma DK, Varshney D, Yadav A. ScienceDirect Detection and veracity analysis of fake news via scrapping and authenticating the web search Action editor : Alessandra Sciutti. Cogn Syst Res 2019;58:217–29. https://doi.org/10.1016/j.cogsys.2019.07.004.

[23] Bhutada S, Lakshmi A, Shravya VM, Reddy JS. REALITY SHOW ANALYTICS FOR TRP RATINGS BASED ON VIEWER ' S OPINION 2018:1327–34.

[24] Roh Y, Heo G, Whang SE. A Survey on Data Collection for Machine Learning: A Big

Data - AI Integration Perspective. IEEE Trans Knowl Data Eng 2019;4347:1–1. https://doi.org/10.1109/tkde.2019.2946162.

[25] Sidhu A, Kanwal E, Attwal PS, Gupta G. To Evaluate & Predict the Television Serials " TRP 2019;6:1–5.

[26] Westerlund O, Bhaumik C, Ghose A, Damratoski KJ, Field AR, Mizell KN, et al. Television goes online: Myths and realities in the contemporary context. J Appl Bus Res 2019;53:225–45. https://doi.org/10.1017/CBO9781107415324.004.

[27] Shayaa S, Jaafar NI, Bahri S, Sulaiman A, Seuk Wai P, Wai Chung Y, et al. Sentiment analysis of big data: Methods, applications, and open challenges. IEEE Access 2018;6:37807–27. https://doi.org/10.1109/ACCESS.2018.2851311.

[28] Zainuddin N, Selamat A. Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st Int Conf Comput Commun Control Technol Proc 2014:333–7. https://doi.org/10.1109/I4CT.2014.6914200.

[29] Potapenko A, Vorontsov K. Robust PLSA performs better than LDA. Eur Conf Inf Retr 2013:806–9. https://doi.org/10.1007/978-3-642-00958-7.

[30] B MTK, Tsarouchis SF. Towards Fashion Recommendation : An AI. vol. 1. Springer International Publishing; 2020. https://doi.org/10.1007/978-3-030-49186-4.

[31] Nguyen TT, Adams N, Huang D, Glymour MM, Allen A, Nguyen QC. State-level racial attitudes and adverse birth outcomes: applying natural language processing to Twitter data to quantify state context for pregnant women (Preprint). JMIR Public Heal Surveill 2019;6:1–12. https://doi.org/10.2196/17103.

[32] Mubarak H, Rashed A, Darwish K, Samih Y, Abdelali A. Arabic Offensive Language on Twitter: Analysis and Experiments 2020.

[33] Robinson MC, Glen RC, Lee AA. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. J Comput Aided Mol Des 2020;34:717–30. https://doi.org/10.1007/s10822-019-00274-0.