Expert Systems With Applications 184 (2021) 115476

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sarcasm Detection using Cognitive Features of Visual Data by Learning Model

Basavaraj N. Hiremath [a,*], Malini M. Patil [b]

[a] Department of Computer Science and Engineering, JSS Academy of Technical Education Bengaluru-560060, Karnataka, India, Visvesvaraya Technological University, Belagavi 590018, Karnataka, India

[b] Department of Information Science and Engineering, JSS Academy of Technical Education Bengaluru, Karnataka, India, Visvesvaraya Technological University, Belagavi 590018, Karnataka, India

ARTICLE INFO

Keywords:
Natural language processing
Multiclass neural network
Sentiment analysis
Sarcasm
Cognitive features
Pragmatic

ABSTRACT

The objective of the paper is to detect sarcasm in human communication. The methodology uses basic cognitive features of human utterances by capturing three modes of data viz., voice, text, and temporal facial features. The captured data is unstructured as it consists of parameters of feelings and emotions to generate sarcasm which affects expressions through glottal and facial organs. The data capturing method is equally challenging as compared to the method of data processing to acquire features. The significant work is aligned to make natural decisions in the prediction processes using cognitive information in the data lineage. Sarcasm detection in natural human communication is a challenging process. The Linguistic features of natural language processing (NLP) methods help identify sentiment as negative and positive sentences based on polarity using the pre-labelled samples. The multiclass neural network model is used as a soft cognition method for the detection of sarcasm under cloud resources. Identified cognitive features have information like voice cues and eye movements, they tend to influence the decision of detecting sarcasm. The visual data are found to be quite interesting and can establish a strong platform in the area of NLP for further research work.

1. Introduction

Sarcasm emanates from natural utterances in human communication. It is a remark made by a person X on person Y during their communication. The remark is always the opposite of one's say either to hurt someone's feelings or to make fun of them (Cambridge Dictionary, 2020). When the utterances involve feelings, the cognitive organs of the human body influence the voice and facial expressions that include eye movement and body language. Utterances are natural outcomes in human communication. They are sentences that carry the feelings which arise out of one's mind. The sentences have features resulting from emotions, linguistics, and psychology that embeds variations from glottal intensities, and facial features at that instance. The automation of human communication can be done by a machine using artificial intelligence, envisage accurate delivery of command, and structure complete form of dialogues. Expression of emotion is intrinsic to humans (Poria, Majumder, Mihalcea, & Hovy, 2019), it is a significant part of artificial intelligence. So, sarcasm depends on the context and discourse of conversation. The research in sentiment analysis of extracting behavioural feelings is vast but not limited to classifying negative and positive sentences using linguistic words (Hiremath, Basavaraj, & Patil, 2019) but also their location and context. The recent advances in natural language processing, learning models with higher accuracy to reproduce and process in various forms of unstructured data like visual, voice, and speech to text forms is the unique approach in the analysis. The cloud technology and use of resources of hardware made a breakthrough in analysis, a step forward to the decision making of behavioural understandings of the communication.

The performance of the NLP system is increased with a graph called jumping NLP curves (Cambria, Poria, Gelbukh, Nacional, & Thelwall, 2017), initially with the evolution of syntactic curve, semantics curve, and then pragmatics curve. The detection of sarcasm aligns with personality recognition and combines human-like performance. Marvin Minsky would say sentiment analysis is one that human minds convey jumbled ideas of emotions and opinions through natural language. The analysis of the manner of speaking and metaphor understanding gives a high-performance solution framework in NLP problems. Artificial intelligence-enabled understanding is knowing conversation, and

* Corresponding author.
E-mail address: basavaraj@ieee.org (B.N. Hiremath).

https://doi.org/10.1016/j.eswa.2021.115476
Received 12 October 2020; Received in revised form 10 May 2021; Accepted 23 June 2021
Available online 26 June 2021

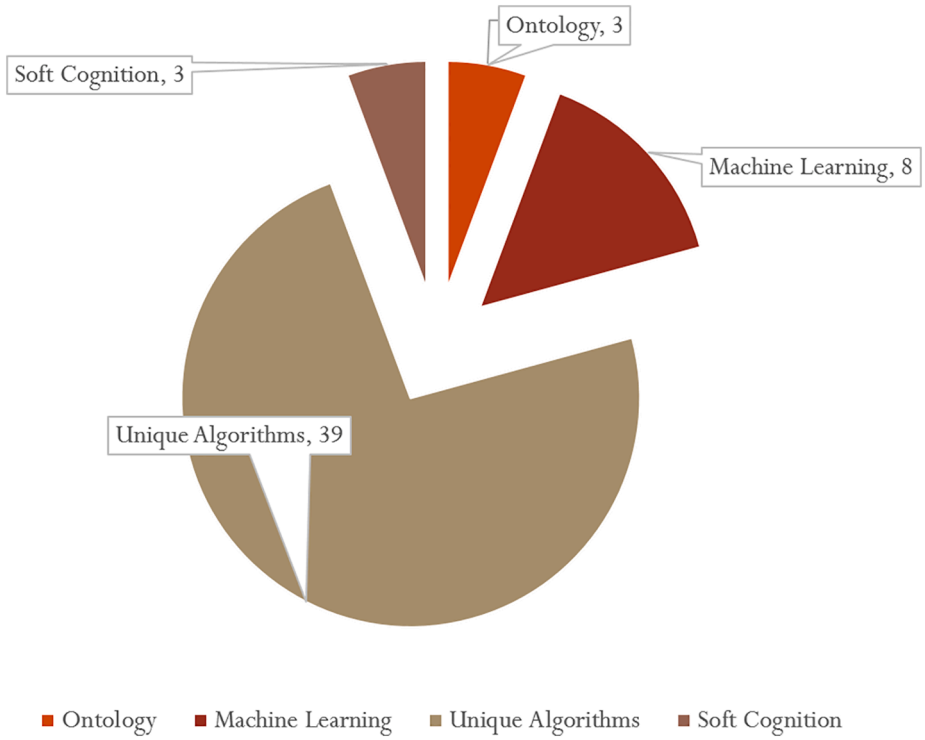

Fig. 1. Graph showing use of the type of algorithms for sarcasm detection in the literature survey.

dialogue systems are the basics of human communication. Sarcasm is expressed through verbal and non-verbal cues viz. tone changes, looking face and syllable drawn-out to overemphasize the word (Castro et al., 2019) that makes automatic classification easier in using a multimodal dataset, called MUS-tARD created out of popular television shows, basically with the context of dialogues. The audio modality performance includes pitch, intonation to find the mean. The visual features are extracted with the pool5 layer of the ImageNet that demonstrates the need for multimodal learning in sarcasm detection. To explore the cognitive NLP paradigm (Mishra, Dey, & Bhattacharyya, 2017), a combination of eye gaze characteristics are acquired along with textual data using tweet snippets, an inference of higher cognitive load on eye movements happens in the utterance of sarcastic remarks because to make pragmatically negative opinions which is complex, subtle eye movement with irregular scan paths leads to higher saccade distance.

The structure of the paper is framed as follows. Section 2 describes the literature reviews related to the proposed research work. Section 3 describes the data capturing methodology and its standards. Section 4 focuses on the methodology of the work. Then discussions of results and conclusion are briefed in Section 5.

2. Literature reviews

A hybrid method designed to detect sentiments with a deep learning model along with a derivative of context-based text analytics and is framed with a comparison of multimodal data in (Kumar, Srinivasan, Cheng, & Zomaya, 2020). An aspect-level sentiment classification for a benchmark dataset fused with context is computed by the authors in the paper (Nguyen & Nguyen, 2020) using deep neural networks. The authors (Bijari, Zare, Kebriaei, & Veisi, 2020) have worked on text mining for semantic and sentence term relationships to increase substantial performance in identifying the sentiment behaviour and are achieved with convoluted network algorithms.

Text analytics is computed in the article (Park, Song, & Shin, 2020) where the bidirectional long short-term memory model on sentences i.e. impact of aspect term on the sentiment classification is done. Learning models used (Liu & Shen, 2020) with a special gated alternate neural network and convolutional neural network to achieve sentiment analysis based on aspect terms. Large acoustic variances (Virtala, Partanen, & Tervaniemi, 2018) which affect the prosodic cues, the speech variability happens with neural discrimination of voice parameters. Research carried out by authors in (Hazarika et al., 2018) with syntactic and context for the lexical analysis is dependent on presumptions, authors point that person-to-person stylistic expressions vary for sarcastic nature, the detection is done with discussion forums with convolutional neural networks. For a dataset of social media (Karoui, Zitoune, & Moriceau, 2017), linguistic analysis is to detect irony for the sentiment analysis and is reviewed for the Arabic text sentences. A multi-modal dataset (Huang, Zhang, Zhang, & Yu, 2017) with emoticon and image in microblogging media was used for sentiment classification latent Dirichlet allocation to identify the hidden topic. A dataset of annotated news videos (Kunneman, Liebrecht, Van Mulken, & Van Den Bosch, 2015; Pereira, Pádua, Pereira, & Benevenuto, 2016) was used in analysing sentiment behaviour with a focus on facial expressions, modulation of speeches, loudness, speech fundamental frequencies and arrive at tension levels of person readers using the neural network by computing facial action coding system i.e. happy, anger, sadness, fear etc.

In (Kunneman et al., 2015) it is assumed that human labelling of sarcasm sentences, twitter datasets are collected to conduct the analysis. It is used to identify the linguistic markers for signalling sarcasm, can be reduced by hashtag with hyperbole and nonverbal expressions. The entire classification is done by machine learning methods. A quantified graph shown in Fig. 1 shows the use of typical algorithms for detecting sarcasm mainly on linguistic analysis as per the literature study reveals.

A neuropsychology study (Matsui et al., 2016) has been done using functional magnetic resonance imaging in context to human behaviour with comprehension to sarcasm. The perception of sarcasm was augmented when positive prosody used in the case of a bad deed, on the contrary, it's not significant in representing good deed when the context is used with negative prosody. A nonlinear SVM (Suhaimin, Hijazi, Alfred, & Coenen, 2017) method is experimented with for the bilingual texts of English and Malay language to take out features related to

Table 1
Details of prominent datasets used in the literature survey.

Reference	Name of dataset	Usability	Summary
(Abdi, Shamsuddin, & Aliguliyev, 2018)	TAC2008, DUC2006, 609 Blog Documents of 25 topics.	Opinion summarization tasks	Opinion on Text Only
(Moussa, Mohamed, & Haggag, 2018)	Opionosise4, TGSum, Hu & Liu Weibo data, DUC (2001–2005) TUC (2010–2011), Gigaword	Topic oriented opinion, Customer review data, Chinese short messaging, Opinion Mining	Social media Text
(Chaturvedi, Ragusa, Gastaldo, Zunino, & Cambria, 2018)	MPQA, MOUD, TASS 2015 Corpus.	For identification of Subjectivity detection	Review dataset With learning model
(Peng, Ma, Li, & Cambria, 2018)	SemEval2014 – Positive, Negative and Multiword selects.	Chinese sentiment analysis	Social media data linguistic analysis
(Xiaomei, Jing, Jianpei, & Hongyu, 2018)	HCR – healthcare reform datasets, OMD – American presidential debate.	To identify weak dependency connections	Microblog datasetss
(Gidhe et al., 2017)	Reddit dataset.	Sarcasm detection on non-tagged statements text only	Text
(Araque, Corcuera-Platas, Sánchez-Rada, & Iglesias, 2017)	SemEval2013 SemEval 2014, Vader, STS-Gold, IMDB, PL04, Sentiment 140.	Sentiment analysis	Semeval database
(Chandra Pandey, Singh Rajpoot, & Saraswat, 2017)	Stanford Twitter Corpus, 1.6 million tweets automatically annotated, Sanders analytical dataset, Twi dataset.	Sentiment analysis	Text of tweets

prosodic, pragmatic, syntax, and lingual are investigated with NLP tools. Sarcasm is directed to the audience, where the utterance is made. The detection of sarcasm is not limited to linguistic and psychological aspects (Cambria et al., 2017), as the works are few in the computational literature. So, the analysis requires a profound understanding of the natural language which results in personality features by developing learning algorithms.

2.1. A brief review on data sets found in the literature

The literature study reveals that the analysis of the sentiment behaviour of human communication requires specific data sets. In sentiment analysis problems, the social media dataset is most frequently used. The datasets varied from using labelled and marked emoticons with sad, angry, happy, and joy remarks. The social media platforms viz. Facebook, Twitter, Instagram have helped to create enormous amounts of dataset repositories. The benchmark datasets refer to laboratory samples, movie-based datasets, newsreaders datasets which are context-based utterances. Table 1 presents a review of prominent data sets used by the researchers. It is found that the text mode of data samples is worked predominantly for sentiment evaluations.

Table 2 depicts the data set collection methods found in the literature. Few multimodal datasets are used for research with context and aspect-based works. But in video analysis, the researchers have aimed to identify facial expressions to recognise 'effect' and 'product review' in the samples. It is also found from Table 2 that Sarcasm detection is also done by gaze fixation for a freely available dataset. Video frames captured by tv shows are sourced out with context-based sarcasm and

Table 2
Details of methods used for data set collection (existing and present).

Existing work			Processes used in the present work
Citation	Name of dataset	Usability	
(Liu, Tang, Wang, & Chen, 2017)	SemEval2014 dataset of restaurant and laptop, Chinese datasets with tweets.	Text data: Aspect based, with gated alternate neural network	Multimodal data Text, Voice cues, and Eye movement and auto transcribed Learning model by Neural network
(Castro et al., 2019)	Popular TV shows dataset (MUStARD), manual transcribed	Multimodal data processing: context-based	
(Filik, Brightman, Gathercole, & Leuthold, 2017)	Read data of participants	Eye movement: Irony Emotional impact by track, Context-based. Hurt response	
(Bishay & Patras, 2017)	Video frames UNBC, FERA Datasets	Video: multi-layer facial expressions Behaviour analysis using CNN and RNN-Affect recognition	
(Chaugule, Abhishek, Vijayakumar, Ramteke, & Koolagudi, 2016)	MPLab GENKI-4K Database	Video: Facial expression for a product review	
(Siritanawan & Kotani, 2016)	Facial images	Video: Facial action units	
(Mishra, Kanojia, Nagar, & Dey, 2016)	Annotated freely available dataset	Video: Sarcasm Naïve Bayes -gaze fixation.	

Table 3
Feature matrix.

Sample_ID	nsa02	nsa1	Sa49
Text_sarcasm_label	No	No	yes
Jitter %	1.927	2.184	2.572
Shimmer %	9.427	8.038	9.833
Median_pitch Hz	136.827	155.657	126.295
Voice_breaks numbers	3	3	11
Total_energy Pascal^2 sec	0.000576	0.043651	0.18875
Max_amplitude_pascal	0.127136	0.580933	0.999969482
Mean_power_db	54.73	74.49	77.41
MeanHarmonics_noise ratio in dB	12.385	13.269	9.583
Eyeball movement numbers	1	0	2
EAR_ratio	0.224018277	0.14889759	0.17043241

eye movement frames were used with gaze fixation, hence the natural utterances were used as unique datasets created to evaluate with different percentage of balancing test and train sets. Thus, it is concluded that the use of unstructured data *namely* voice and video has prompted *the direct use of natural utterances into the communication media*.

From Table 2, it is clear that in the literature survey the authors have used the multimodal data as per their objective. The last column justifies the usage of multimodal data in the present work.

3. Methodology

The objective of the proposed paper is to detect sarcasm on cognitive features of the multimodal data which includes voice, video and text data. The methodology is presented in two phases.

Phase 1: Transformation and processing of multimodal data.

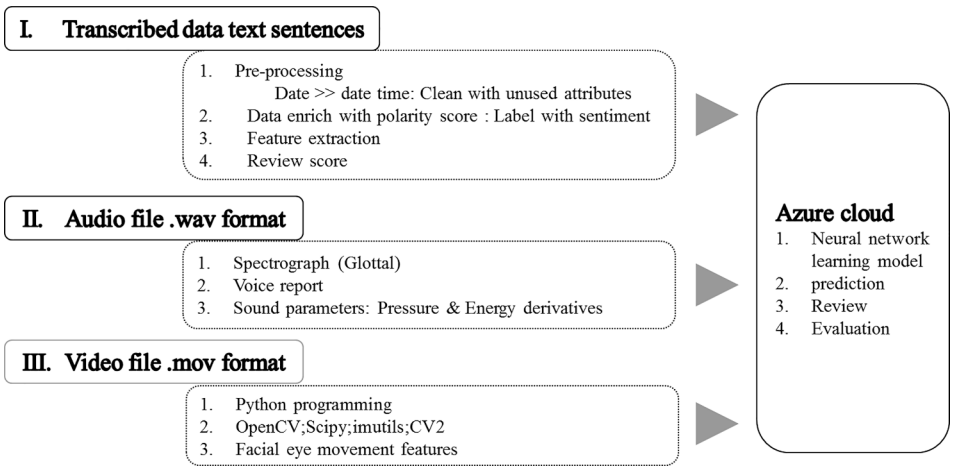

Fig. 2. Experimental set up of transformation and processing of integrated data.

- Phase 1 is about the transformation and processing of integrated data as depicted in Fig. 2.
- Capturing significant representative samples of data including text, voice, and video that arise out of natural utterances of sentences in English as human communication.
- The text, voice and video data are transformed and processed in this step after capturing.
- Pre-process the final acquired features in the form of a feature matrix (*.csv* file) with categorical, numerical and decimal numbering format.
- Cloud architecture in Microsoft Azure (Research, 2020) is selected for the advantage of scalable, platform-independent, and ease of usage of hardware and software resources with security.
- A cluster of platforms is framed to classify the cognitive features using artificial intelligence and machine learning framework

A detailed explanation of the data capturing process for text, video and voice is presented below.

(i) Text Data:

The text analytics is carried out with the programming language Python using the TEXT BLOB package for the parser and identifying the polarity to detect linguistic features (Hiremath & Patil, 2020; Bharti, Vachha, Pradhan, Babu, & Jena, 2016) called as a negative sentence with positive situation and negative situation with the positive sentence. This labelled information is considered as a linguistic feature to the next step with unique sample identification.

(ii) Voice data

Voice analytics is carried out as parameters of the spectrograph. After capturing the *.wav* file of the sample for processing to extract the feature of decision making to achieve the objective of finding cognitive information emanated from the utterance naturally from glottal parameters. These voice cues are processed by using the basic rule of physics to revolve around the air pressure articulated from the mouth. In Fig. 2, the process flow is presented to extract voice cues. Jitter (Local) is the measure of voice quality in percentage, which is derived by Eq. (1).

$$\text{Jitter} = \sum_{i=2} N|T_i - T_{i-1}|/(N-1) \quad (1)$$

where T is the duration of *Ith* interval in N intervals. Shimmer and the harmonics to noise ratio which are derivatives of the energy of sound are captured as voice cues. The open-source tool Praat (van Heuven & Boersma, 2001) is used to accurately extract the voice cues using voice reports and sound information as the preliminary derivatives. The total energy in pascal and mean power intensity in the air(decibels) are measured and taken for processing. A detailed case study is carried out in extracting parameters related to glottal intensities, a measure of intensity and pressure are presented in Hiremath et al. (2019).

(iii) Video data

Video analytics is processed with the video sample in *.mov* format to analyse the facial features mainly the eye aspect ratio as per the framework (Adrian rosebrock, 2017). The blinks and the eye aspect ratio are calculated as a cognitive feature that arises at the time of the utterance of sentences by human communication. The features are significant to impact the prediction of results i.e., sarcasm or not sarcasm. Fig. 2 represents the programs, packages used for computation. The facial eye movement prompts are independent of human communication at the instance of the utterance of complex opinions. The behavioural information is captured automatically and delegates the same to the learning system.

The classification learning model is used to predict the sarcasm label as true or false.

Phase II: Learning Model.

Phase II is about creating a learning model using a multiclass neural network for the prediction of sarcasm using multimodal features acquired from voice cues, visual features, and text labels and is shown in Fig. 3.

The steps are briefed as follows:

1. Pre-processing: The pre-processing of transcribed text data is explained in phase 1.

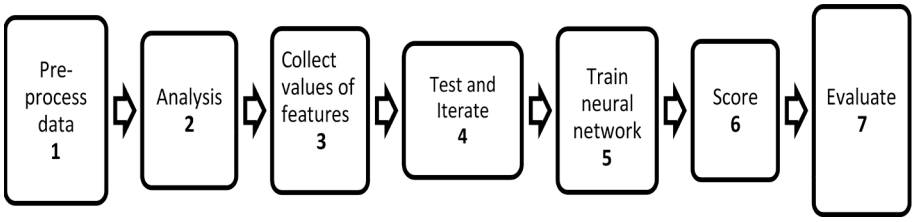

Fig. 3. Process flow diagram for learning model using multiclass neural network.

Table 4
The score model tabulation of 75/25 split ratio for training and test data set.

Sample_ID	Text sarcasm Label	Jitter	Shimmer	Median_Pitch	Voice_breaks	Total_energy	Max_Amplitude_pascals	Mean_power_db	Hormonics_Noise	Eye ball movement	Ear_ratiop	Scored probabilities for class "no"	Scored probabilities for class "yes"	Scored labels
SU_MA_04	yes	2.042	16.357	113.36	8	0.001235	0.133118	60.11	7.632	1	0.161126	0.021844	0.978046	no
SA_FE_04	Yes	1.787	11.706	270.157	11	0.011508	0.343903	66.87	9.072	1	0.236735	0.001851	0.998147	yes
nsa02	no	1.927	9.427	136.827	3	0.000576	0.127136	54.73	12.385	1	0.224018	0.999308	0.000688	no
Sa60	yes	2.671	11.402	101.514	11	0.026194	0.662323	70.25	8.732	1	0.128799	0.188522	0.811244	yes
SA_MA_04	yes	2.505	6.78	118.555	7	0.192108	0.999969	79.63	8.347	2	0.249991	0.000005	0.999995	yes
Sa78	yes	2.386	10.202	145.804	8	0.041861	0.607239	72.24	11.663	3	0.116411	0.261571	0.737895	yes
Sa77	yes	3.195	12.737	105.396	12	0.024297	0.580475	68.4	8.876	1	0.139831	0.720584	0.279438	no
SU_MA_05	yes	2.254	15.405	136.55	6	0.005074	0.358124	65.73	7.939	5	0.281449	0.006265	0.993731	yes
Sa70	yes	2.807	13.438	116.725	5	0.040482	0.999969	73.77	9.98	1	0.132439	0.000926	0.999076	yes
Sa64	yes	2.183	11.342	117.751	15	0.101125	0.842133	75.61	11.059	0	0.126784	0.009029	0.990967	yes
Sa61	yes	2.342	11.743	109.933	6	0.040197	0.847412	73.06	10.23	0	0.183557	0.002519	0.997481	yes
NSA_50	no	2.54	12.38	112.971	13	0.030852	0.4617	67.41	9.853	4	0.165653	0.559181	0.4403	no
NSA_53	no	2.456	13.103	111.817	21	0.027	0.49231	66.26	9.443	3	0.17912	0.408813	0.590945	yes
SA_MA_05	Yes	1.878	9.312	138.974	15	0.218478	0.999969	77.26	10.06	1	0.21523	0.000008	0.999992	yes

2. Analysis: The analysis of labelling and data validation with their units from respective outputs is carried out.
3. Values of features: Feature values are collected from all datasets which influence the decision of categorising Sarcasm utterance. The significant feature extraction from all datasets of Phase I has tabulated accordingly with a unique sample ID.

   Feature capture:

   The cognitive features of the multimodal data are: Text, glottal signatures in case of voice data, eyeball movement in video data respectively. The authors of the paper have already conducted independent experiments on text and glottal signatures as presented in section 3 of the paper. These features are informative, non-redundant and influential for predictive parameter and are basically decisive in nature for pointing as knowledge.
4. Test and Iterate: The testing set is grouped by iterating the process to decide the quantification of samples.
5. Train the neural network: The classification with both the test and train sets to be passed in the neural network model for a prediction.
6. Record the Score: After validation of the percentage of confidence, results are checked to record the score.
7. Evaluate: The statistical evaluation is done with the standard practice. As per the scored labels, mentioned in Table 4 in section 4, the occurrence of probabilities of two classes "no" and "yes" are tabulated. When compared to the input data source with a known labelled class, a percentage is derived out of the standard confusion matrix mentioned in section 4 of (ii).

4. Experiment and results

The data is captured with the standard hardware specifications with low noise in the data collection as per the flow diagram shown in Fig. 4. The respective format of the text, audio file, and video file is .txt, .wav and .mov. The processing of data is done with standard labelling and pre-processing.

The detailed data capturing process for voice and video, with necessary software and hardware configurations is presented below.

(i) Voice and video capturing hardware specifications of samples:

Lavalier condenser microphones are used with a laptop camera as an integrated device, in Omnidirectional. The attachment of the microphone is upside down to reduce extra breath noises and tonal inconsistencies with head movements at the time of utterance. A foam windscreen is used to remove any external wind noise. The signal to noise ratio and output impedance is as per the standard hardware specifications mentioned as follows.

a. *Frequency Range 65 Hz –18 Hz*
b. *Signal/Noise 74dBSPL*
c. *Sensitivity 30 dB ± 3 dB*
d. *Output impedance 1000 Ohm or less*

(ii) Video capture environment specification

The video is captured in .mov format wherein the memory size of each sample varies from 3.0 Mb to 11.0 Mb. The output of the multimedia container file is created by recording in normal desktop lighting conditions and laboratory air flow, visual capture window is adjusted to capture high resolution, high fidelity data file which can store video, audio, and also text. This file is also called a quick time movie file, which is produced by the MAC apple operating software with the following specifications.

a. *MAC OS – Mojave 10.14.6*
b. *HD Graphics 6000–1536 MB: Intel core i5 1.8 GHz*
c. *Camera application used: photo booth*
d. *Audio capture environment specification*
e. *Output .wav file (wave file format) the memory size of each sample varies from 1.0 MB to 2.0 Mb.*

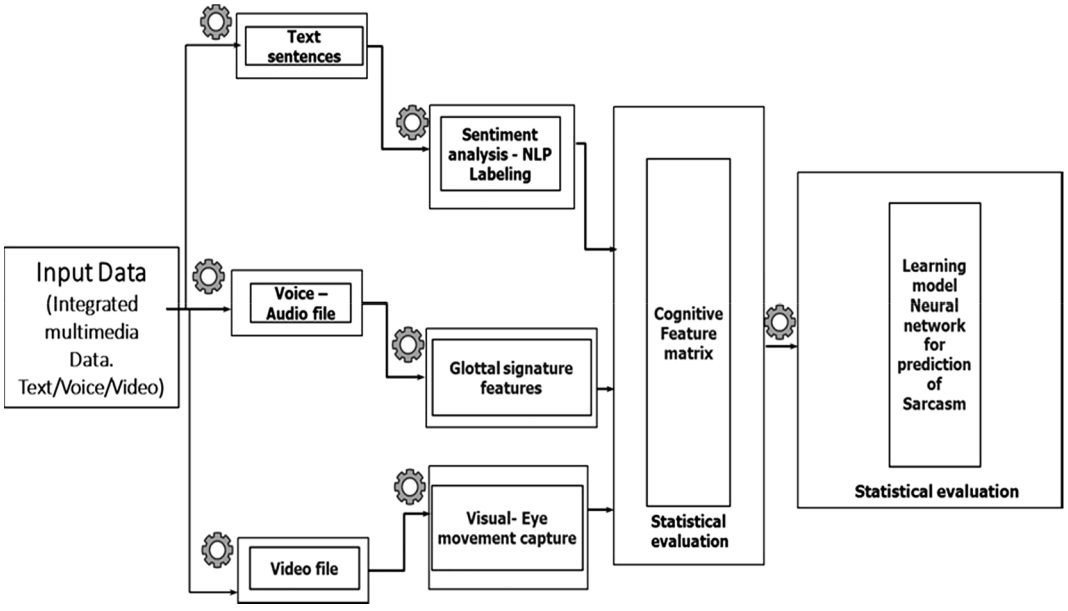

Fig. 4. The flow of the data capturing process.

f. *A sampling rate 44100 kHz @ CD-quality [online open source]*

Once the data capturing is done the steps for creating a learning model are explained below.

- The experiment is processed with one unique entity of video file which comprises all the information to maintain the integrity of the sample. It is carried out with an individual video file to maintain data lineage after transformed into respective audio and text file.
- The video file is created and captured with a hi-fidelity audio microphone, then to maintain data lineage through transformation it is processed as a single entity file.
- As mentioned, the cognitive features in voice cues and video frames are carried forward to create an array of feature matrix. As these are cognitive patterns, which cannot be anticipated for possible situations. The nature of these patterns is dynamic with respect to sample and time, hence, to avoid probable human errors, creating a learning model is found to be an appropriate option.
- The experiment is conducted on the captured data set, in the standard laboratory environment. The comparisons are done by varying the training and test data quantities within the representative samples. The captured datasets are unique and representative by category as utterances are generated by male (adult), female (adult) voices.
- The transcribed data after the transformation from the audio file will be processed as a text-linguistic sentence for processing.
- The users are referring to this, for identifying significant features from the dataset captured. To capture natural utterances, the illustrative samples of video mode of datasets in the latest compressed format using an ecosystem of hardware having lesser noise at the capture point.
- The sample is collected by the natural utterance of sentences and saved as a video file with detailed specifications of the data description. The specifications are highlighted in the video and audio capture details with high fidelity and high resolution, each data source file is in .mov format. The size varies from 3.0 megabytes to 11.0 megabytes in size depending upon the length of sentence utterance. All the specification of selecting a representative sample is taken care of.
- Significant features are saved as matrix and processed for classification in Microsoft azure platform, a multiclass neural network learning model is framed for classification as shown in Fig. 5 with one hidden layer.

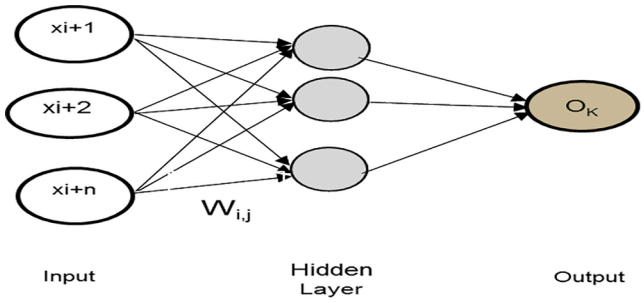

Fig. 5. The featured matrix consisting of text, visual, and voice cues are tabulated in Table 3.

The type of normalizer selected is by the min–max method, the significant characteristics are that it preserves the relationships among the actual values of data. Eq. (2) describes the value of the computation.

$$vi = \frac{vi - minA}{maxA - minA}(new - maxA - new_{minA}) + new\_minA \tag{2}$$

The neural network is connected where input and output units are associated with weights. The learning process happens by adjusting weights from the classified labels of input tuples.

The confusion matrix for classification computes to the overall accuracy of 78.5714%, then true positive accuracy of 81.8%. Fig. 6 displays the minimization of mean error over iterations, the estimated pre-training mean error was 1.88, the post-training mean error was concluded at 0.000464.

Table 4 displays the feature matrix values along with the probability score for Sarcasm 'yes and sarcasm 'no' with the scored labels. The score model tabulation of 75/25 split ratio for training and test data set. The confusion matrix for classification computes to the overall accuracy of 78.5714%., then true positive accuracy of 81.8%.

5. Conclusion

As stipulated the drill down fact for sarcasm behaviour in the natural

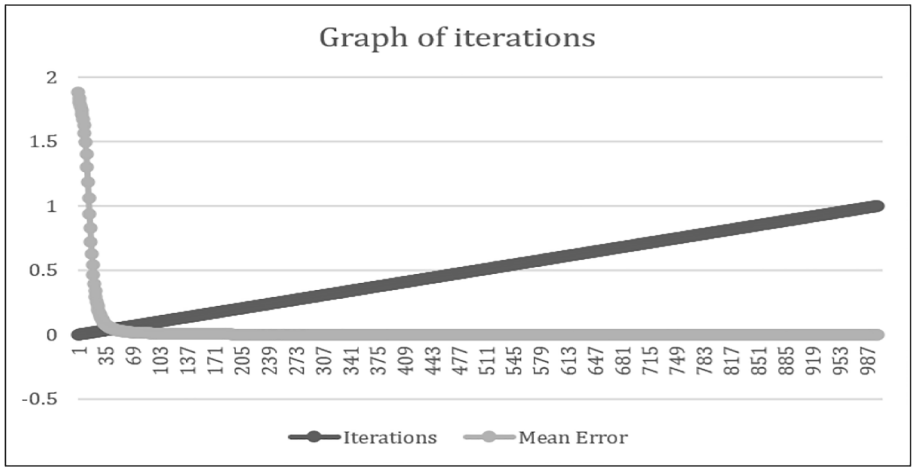

Fig. 6. Graph showing the curve for iterations v/s mean error.

utterances is challenging to extract the knowledge with a decision. In this proposed research experiment it is framed to capture the cognitive features that emanate at the time of utterance from human communication to understand the influence of each feature in deciding the sarcasm or no sarcasm sentence. The outcome of the literature survey in section 2, points to ample usage of text analytics i.e. linguistic where the dependence is on the use of word meaning, affect, and aspect of sentences. The effort of using a neural network as a soft cognition tool to classify and predict the sarcasm from the utterances with the use of voice clues and facial features gave knowledge from the information acquired from three modes of datasets. Though the datasets are unstructured, the programming language Python and its OpenCV packages and the use of open-source Praat have helped to build a unique framework with statistical accuracy checks.

Future Scope: The scope of future work is to conduct experiments on typical conversations with context-based, dialogues and benchmark datasets to be able to get fair learning performance and test them on language-independent utterances to prove the impact of pragmatic/cognitive features in human behaviour while uttering sentences.

CRediT authorship contribution statement

Basavaraj N. Hiremath: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Malini M. Patil:** Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Authors acknowledge their organisation J S S Academy of Technical Education Bengaluru, for providing the support in the research work.

References

Abdi, A., Shamsuddin, S. M., & Aliguliyev, R. M. (2018). QMOS: Query-based multi-documents opinion-oriented summarization. *Information Processing and Management, 54*(2), 318–338. https://doi.org/10.1016/j.ipm.2017.12.002

Adrian Rosebrock. (2017). Facial landmark predictors. Retrieved February 20, 2020, from https://www.pyimagesearch.com/category/facial-landmarks/.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications, 77*, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002

Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: A big data approach. *Digital Communications and Networks, 2*(3), 108–121. https://doi.org/10.1016/j.dcan.2016.06.002

Bijari, K., Zare, H., Kebriaei, E., & Veisi, H. (2020). Leveraging deep graph-based text representation for sentiment polarity applications. *Expert Systems with Applications, 144*. https://doi.org/10.1016/j.eswa.2019.113090

Bishay, M., & Patras, I. (2017). Fusing Multilabel Deep Networks for Facial Action Unit Detection. Proceedings – 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 – 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge, 681–688. https://doi.org/10.1109/FG.2017.86.

Cambria, E., Poria, S., Gelbukh, A., Nacional, I. P., & Thelwall, M. (2017). Affective Computing And Sentiment Analysis Sentiment Analysis Is a Big Suitcase.

Cambridge Dictionary. (2020). Retrieved August 22, 2020, from https://dictionary.cambridge.org/dictionary/english/sarcasm.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards Multimodal Sarcasm Detection (An_Obviously_ Perfect Paper), 4619–4629. https://doi.org/10.18653/v1/p19-1455.

Chandra Pandey, A., Singh Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing and Management, 53*(4), 764–779. https://doi.org/10.1016/j.ipm.2017.02.004

Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., & Cambria, E. (2018). Bayesian network based extreme learning machine for subjectivity detection. *Journal of the Franklin Institute, 355*(4), 1780–1797. https://doi.org/10.1016/j.jfranklin.2017.06.007

Chaugule, V., Abhishek, D., Vijayakumar, A., Ramteke, P. B., & Koolagudi, S. G. (2016). Product review based on optimized facial expression detection. *IEEE*.

Filik, R., Brightman, E., Gathercole, C., & Leuthold, H. (2017). The emotional impact of verbal irony: Eye-tracking evidence for a two-stage process. *Journal of Memory and Language, 93*, 193–202. https://doi.org/10.1016/j.jml.2016.09.006

Gidhe, P., Suhaimin, M. S. M., Hijazi, M. H. A., Alfred, R., Coenen, F., Prasad, A. G., … Wang, X. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications, 53*(4), 467–472. https://doi.org/10.1016/j.eswa.2017.02.002

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., & Mihalcea, R. (2018). CASCADE: Contextual Sarcasm Detection in Online Discussion Forums, 1837–1848. Retrieved from http://arxiv.org/abs/1805.06413.

Hiremath, B. N., & Patil, M. M. (2020). Enhancing optimized personalized therapy in clinical decision support system using natural language processing. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2020.03.006

Hiremath, B. N., & Patil, M. M. (2019). Analysis of voice cues in recognition of sarcasm. *Recent Patents on Computer Science, 12*, 1–15. https://doi.org/10.2174/2213275912666190819113541

Hiremath, Basavaraj N, & Patil, M. M. (2019). Analysis of speech in human communication. Journal of Computer Science and Software Testing, 5(2), 8–16. https://doi.org/http://doi.org/10.5281/zenodo.3250518 Abstract.

Huang, F., Zhang, S., Zhang, J., & Yu, G. (2017). Multimodal learning for topic sentiment analysis in microblogging. *Neurocomputing, 253*, 144–153. https://doi.org/10.1016/j.neucom.2016.10.086

Karoui, J., Zitoune, F. B., & Moriceau, V. (2017). SOUKHRIA: Towards an irony detection system for arabic in social media. *Procedia Computer Science, 117*, 161–168. https://doi.org/10.1016/j.procs.2017.10.105

Kumar, A., Srinivasan, K., Cheng, W. H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing and Management, 57*(1). https://doi.org/10.1016/j.ipm.2019.102141

Kunneman, F., Liebrecht, C., Van Mulken, M., & Van Den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management, 51*(4), 500–509. https://doi.org/10.1016/j.ipm.2014.07.006

Liu, N., & Shen, B. (2020). Aspect-based sentiment analysis with gated alternate neural network. *Knowledge-Based Systems, 188*, Article 105010. https://doi.org/10.1016/j.knosys.2019.105010

Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics, 75*, S34–S42. https://doi.org/10.1016/j.jbi.2017.05.023

Matsui, T., Nakamura, T., Utsumi, A., Sasaki, A. T., Koike, T., Yoshida, Y., … Sadato, N. (2016). The role of prosody and context in sarcasm comprehension: Behavioral and fMRI evidence. *Neuropsychologia, 87*, 74–84. https://doi.org/10.1016/j.neuropsychologia.2016.04.031

Mishra, A., Dey, K., & Bhattacharyya, P. (2017). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. ACL 2017 – 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1, 377–387. https://doi.org/10.18653/v1/P17-1035.

Mishra, A., Kanojia, D., Nagar, S., & Dey, K. (2016). Harnessing Cognitive Features for Sarcasm Detection. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 1095–1104.

Moussa, M. E., Mohamed, E. H., & Haggag, M. H. (2018). A survey on opinion summarization techniques for social media. Future Computing and Informatics Journal, (2017). https://doi.org/10.1016/j.fcij.2017.12.002.

Nguyen, H. T., & Nguyen, L. M. (2020). ILWAANet: An Interactive Lexicon-Aware Word-Aspect Attention Network for aspect-level sentiment classification on social networking. *Expert Systems with Applications, 146*, Article 113065. https://doi.org/10.1016/j.eswa.2019.113065

Park, H. J., Song, M., & Shin, K. S. (2020). Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories. *Knowledge-Based Systems, 187*, Article 104825. https://doi.org/10.1016/j.knosys.2019.06.033

Peng, H., Ma, Y., Li, Y., & Cambria, E. (2018). Learning multi-grained aspect target sequence for Chinese sentiment analysis. *Knowledge-Based Systems, 148*, 55–65. https://doi.org/10.1016/j.knosys.2018.02.034

Pereira, M. H. R., Pádua, F. L. C., Pereira, A. C. M., Benevenuto, F., & Dalip, D. H. (2016). Fusing Audio, textual and visual features for sentiment analysis of news videos, (2015). Retrieved from http://arxiv.org/abs/1604.02612.

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access, 7*, 100943–100953. https://doi.org/10.1109/access.2019.2929050

Research, M. (2020). Azure machine learning report. Retrieved from https://docs.microsoft.com/en-us/archive/blogs/machinelearning/neural-nets-in-azure-ml-introduction-to-net.

Siritanawan, P., & Kotani, K. (2016). Facial action units detection by robust temporal features. In *Proceedings of the 2015 7th International Conference of Soft Computing and Pattern Recognition*. https://doi.org/10.1109/SOCPAR.2015.7492801

Suhaimin, M. S. M., Hijazi, M. H. A., Alfred, R., & Coenen, F. (2017). Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts. In *ICIT 2017–8th International Conference on Information Technology* (pp. 703–709). https://doi.org/10.1109/ICITECH.2017.8079931

van Heuven, V., & Boersma, P. (2001). Speak and unSpeak with PRAAT. *Glot International*.

Virtala, P., Partanen, E., Tervaniemi, M., & Kujala, T. (2018). Neural discrimination of speech sound changes in a variable context occurs irrespective of attention and explicit awareness. Biological Psychology, 132(October 2017), 217–227. https://doi.org/10.1016/j.biopsycho.2018.01.002.

Xiaomei, Z., Jing, Y., Jianpei, Z., & Hongyu, H. (2018). Microblog sentiment analysis with weak dependency connections. *Knowledge-Based Systems, 142*, 170–180. https://doi.org/10.1016/j.knosys.2017.11.035