SSRG International Journal of Computer Science and Engineering
ISSN: 2348 – 8387 /doi:10.14445/23488387/IJCSE-V8I12P101

Volume 8 Issue 12, 1-5, Dec 2021

Detection of Sarcasm through Tone Analysis on video and Audio files: A Comparative Study On Ai Models Performance

Ayush Jain[1], Prathamesh Patil[2], Ganesh Masud[3], Prof. Sunantha Krishnan[4], Prof. Vijaya Bharathi Jagan[5]

[1]*Information Technology, DBIT, University of Mumbai, India*
[2]*Information Technology, DBIT, University of Mumbai, India*
[3]*Information Technology, DBIT, University of Mumbai, India*
[4]*Information Technology DBIT, University of Mumbai, India*
[5]*Information Technology, DBIT, University of Mumbai, India*

Received Date: 07 November 2021
Revised Date: 11 December 2021
Accepted Date: 23 December 2021

Abstract *- During the past few years, there has been a lot of increase in interest in the field of Sentiment Analysis. Sentiment Analysis is used to analyze a given data and help us to understand the sentiment behind the multimedia data, namely text, audio, and video. Unstructured Big Data has its own challenges, and Detection of Sarcasm is one of the major challenges in it. Sarcasm normally signifies the opposite in order to mock or convey contempt, a definition in the Oxford dictionary. Although there has been a lot of research on sarcasm, most of it is on text data, and very few are over audio and video data. Many times, the word or a sentence may not be sarcastic but are spoken sarcastically with a change in the tone or in pitch. In this system, we propose a mechanism to detect sarcasm in the speech by analyzing the audio using pitch frequency, the stress in the pronunciation as a major parameter as an input to the CNN, LSTM, and Bi-Directional LSTM models for audio recognition and classification. This system can be used in social media websites like Twitter, Facebook, Instagram, and YouTube to help users to identify non-sarcastic videos or audios before actually playing or listening to them and wasting their bandwidth.*

Keywords *- Sarcasm, Audio mining, Speech recognition, MFCC, Contempt.*

I. INTRODUCTION

Communication is simply the act of transferring information from one place, person, or group to another to convey a certain message [2]. Our voice does more than combine sounds to form words. It conveys your mood, emotion, and perspective. The Tone of one's voice is one's ability to change the meaning of the words you convey by changing one's pitch, intonation, volume, and tempo. Because listeners use sound to interpret your message, being sensitive to how one's tone of voice affects what they hear can make one a better communicator [3]. Sarcasm is one of the emotions that can be conveyed by verbal communication and by changing the tone of a voice. Sarcasm is a form of verbal irony that is used to mock or convey something. It is an ironic remark which is sometimes tempered with humor. It is also a use of words that mean the opposite of what you and I really want to say. Example: "Brains are not everything. In fact, in our case, they are nothing…". Sometimes a sentence in text format is not sarcastic but is said with a sarcastic tone by the change in pitch or volume, or tempo. E.g., "I 'm really looking forward to being seated for all those exams in a few weeks 'time". Now here, this person wants to say the opposite of what he said using a sarcastic tone by changing his pitch and tempo, but if this wasn't 't said with a particular tone, then the receiver would misinterpret what the sender meant to say. Here figures of speech and tone analysis come into the identification of sarcasm. Identification of sarcasm using tone as a major.

The parameter is one of the difficult tasks in Natural Language Processing (NLP). This project uses a Dataset created from scratch consisting of balanced sarcastic and non-sarcastic clips from various Indian Standup comedies with 2 main Indian accents (South Indian and North Indian). After the data set creation, the video clips and Converted into Waveform Audio File Format (. WAVE or . WAV*)* files using an open-source MP4 to WAV file converter. Audio features are extracted using Librosa Library, where there are a number of extractions features available. The preprocessed and extracted features are then used by the Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi- LSTM), and Convolutional Neural Network (CNN) for classification and comparative analysis of these models. What is the current scenario on sarcasm detection? What is the need of the proposed system(current research, gap that lead to the need of the system, set of objectives, importance of the)

II. RELATED WORK

In recent years, sarcasm detection has gained much popularity in human-machine interaction research and sentiment analysis. Various research projects have approached this problem statement with different sets of implementations through text, audio, video data streams. Although there is limited research on audio and video data streams. For understanding sarcasm detection through audio and video data streams, we referred to two papers

1. Understanding Sarcasm in speech Through Mel-frequency Cepstral Coefficient: Researchers have created their own dataset and also implemented MPEG, HMM, ASR algorithms. They have used MFCC to calculate the intensity of pitch which helped them to acquire accuracy of 70%.
2. Towards Multimodal Sarcasm Detection: Researchers have implemented SVM algorithms also created the MUSTARD dataset. They have carried out multiple evaluations where they found out that multimodal variants significantly outperformed unimodal counterparts with a relative error reduction of 12.9%

III. METHOD FOR DATA COLLECTION, EXTRACTION, AND TRANSFORMATION

To discover the operation of detecting sarcasm through audio tone, we introduce a new dataset consisting of short videos manually annotated for their sarcasm property.

A. Data Collection

To collect potentially sarcastic examples, the team did a web scraping of YouTube videos of our choice. An AI-powered web scraping tool was used for web scraping these YouTube videos for the keywords such as Indian standup comedy, standups, Rahul Subramanian, Anirudh, Anirban Dasgupta, etc.

Web scraping strategy was used to obtain videos from Stand-Up comedy shows. These videos were of average duration of 10-20 minutes. We further used an open-source video editing software, which enabled us to create 5-25 second clips containing sarcastic and non-sarcastic .MP4 files. Our dataset contains 471 sarcastic and 206 nonsarcastic clips.
As a result, 57.3% of the clips from the standup dataset are single sentences, while the remaining clips consist of two or more sentences.

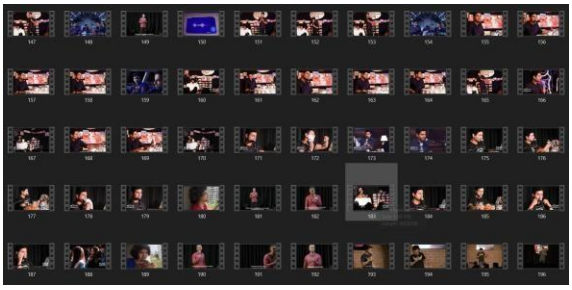

Fig. 1. Sample Dataset

The sarcasm dataset, under a stored folder, consists of videos from three main TV shows: Friends, The Golden Girls, and Sarcasm Alcoholics Anonymous. Initially, they gathered data by clipping the videos, where utterance and context were the considered parameter. Later they developed a web portal where videos were uploaded for a manual labeling process by collecting responses from real-time users [2]. This dataset comprises 690 videos with an even number of sarcastic and non-sarcastic labels. As a result, 61.3% of the utterances from the dataset are single sentences, while the remaining utterances consist of two or more sentences.

We used an open-source editing tool to manually cut down the original 10-20 min video to 5-25 seconds clips. The clip's size varies upon the size of the sentence spoken. The team managed to trim a total of 677 clips, fully spoken sentences from various standup videos.

Steps followed to clip the videos.
There are a total of 3 steps used to clip the videos i. Imported MP4 file into the open-source software ii. Set the start and end parameter to clip the required part of the video iii. Export the clip into the desired folder.

B. Annotation Process

The system interface for labeling the clips manually to the system is as shown below.

The web annotation interface was built, where each video is played individually, and users are allowed to manually give a response whether they felt the video is sarcastic or not. These videos were stored in a database. CSV file format, which was then used for training and testing the model.

IV. Video to Audio data Transformation The reason for choosing the. WAV file format

WAV is an uncompressed format that means the recording is reproduced without any loss in audio quality.
WAV is a simple format file that is fairly easy to process and edit. Other lossy formats exploit general human hearing to reduce file size. That was the only reason for it to be used, thus causing quality loss. Perceptible hearing depends on the user and the amount of compression used.

V. Data Analysis

The experiment was conducted using three different baseline models, namely CNN, LSTM, and Bi-LSTM.
Hyperparameter tuning was used for all three models.

A. Feature Extraction

Mel Frequency Cepstral Coefficient (MFCC) is used to identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion, etc. Any sound generated by humans is determined

by the shape of their vocal tract (including the tongue, teeth, etc.). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the temporal power spectrum of the speech signal is representative of the vocal tract, and MFCC (which is nothing but the coefficients that make up the Mel-frequency cepstrum) accurately represents this envelope. While considering MFCC, we experienced variance in its value and the reason since our dataset consisted of clips of varied duration. We furthered the calculated mean, standard deviation, and difference between MFCC'S feature value. As our dataset is not balanced, we performed under sampling and oversampling.

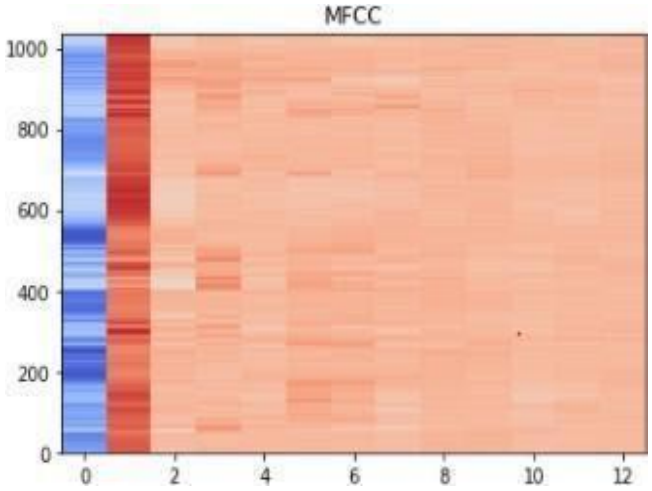

Fig. 2. Signal representation into MFCC format

Generally, the first 13 coefficients (the lower dimensions) of MFCC are taken as features as they represent the envelope of spectra. And the discarded higher dimensions express the spectral details. For different phonemes, envelopes are enough to represent the difference, so we can recognize phonemes through MFCC.

B. Classification using CNN, LSTM, and Bi-lstm

The extracted features are fed into the three models, and then models are trained to classify the audio data as sarcastic or not.

C. Data Augmentation

The approach of synthesizing new data from the available data is referred to as 'Data Augmentation'. It can be used on a small scale as well as large scale data which helps models to perform better on a dataset. As we had just created a handful of videos, we decided to use data augmentation techniques on our dataset. Different methods like shrinking, stretching, noise, shifting, etc., were applied on a dataset of about 677. For every data augmentation technique, parameters values are selected by trying various values and selecting only those which have provided a different sound than the original sound. Additionally, the dataset was updated with data which was modified using data augmentation parameters like

- Noise
- Increase pitch
- Decrease pitch
- Speed and pitch
- Slow speed
- Fast speed

Validation and comparison of the model performances on Sample 2 dataset for the performance measured parameters.

Table 1. System performance on the sample 2 dataset

Experiments	Precision	Recall	F1-score	Accuracy
LSTM	80.8	80.59	80.69	80.54
CNN	70.99	70.56	70.77	70.54
Bi-LSTM	56.38	48.67	52.24	48.62

Table 2. System performance on the sample 1 dataset

Experiments	Precision	Recall	F1-score	Accuracy
LSTM	88.71	88.81	88.76	88.8
CNN	73	73.28	73	73.11
Bi-LSTM	66	100	80	66

Table 3. System performance on the merged dataset (sample 1 and sample 2)

Experiments	Precision	Recall	F1-score	Accuracy
LSTM	89.65	89.65	89.65	89.65
CNN	75.96	75.87	75.88	75.68

Table 4. System performance on the sample 2 dataset by using complete MFCC features

Experiments	Label	Precision	Recall	F1-score	MSE	MAE	Accuracy
CNN (without data augmentation)	0	42	46	44	28	42	57
	1	67	63	65	28	42	57
CNN (with data augmentation)	0	86	96	91	7	10	90
	1	95	85	90	7	10	90
CNN (oversampling + data augmentation)	0	74	70	72	13	18	82
	1	86	88	87	13	18	82
CNN (under-sampling + data augmentation)	0	78	79	78	17	23	79
	1	79	78	79	17	23	79

VI. Results and Conclusion

Performance of three systems upon testing and performance of the system upon real-time prediction While experimenting with CNN, LSTM, and Bi-directional LSTM, what we observed is, CNN gives a better accuracy on training and testing both MUSTARD and Stand Comedy dataset. LSTM was slightly less accurate than CNN, but while predicting the real-time data, LSTM was more accurate than CNN. Also, we performed prediction over real-time data with data augmentation. Here too, LSTM outperformed CNN and Bi-directional LSTM, overtaking the mean of MFCC features and PCA of MFCC but LSTM got overfit on full MFCC feature. With Bi-directional LSTM, it was accurate with the prediction for 60% of the time, which is also similar in the case of CNN's prediction accuracy, while LSTM outperformed both the models with a score of 7-8%. Furthermore, the dataset can be extended to Terabyte, which in all improves the learning accuracy of the model. Also, with the help of the domain expert model, accuracy can be increased; likewise, an error can be improved. Using cloud CPU Instance can be used to derive results in an optimized time. Dataset can also be used with a transformer with different parameters for better results. Also, with image recognition, better results can be observed. CNN model without data augmentation on using full MFCC feature didn't provide us with desired results, so after performing data augmentation, we achieved 90% accuracy with an MSE error rate of 7% and MAE error rate of 10%, and F1- score of 0.90. We had an imbalanced dataset, so we carried out under sampling and oversampling process. On oversampling, we achieved an accuracy of 82%, F1- score of 0.87, MSE of 13%, and MAE of 18%. As we were referenced [1] and [2] papers for multi-modal detection of sarcasm, we found out the model approaches used in both the papers for identifying sarcasm were supervised machine learning algorithms. So, we decided to use an unsupervised machine learning algorithms approach for this paper. [1] and [2] these papers created their own dataset, which resulted in distinctive results achieved.

We also created our own dataset as we were unable to gather a good amount of data from any open-source medium. Also, we picked the dataset created by the [2] and performed the 3 subsequent types of research (i.e., Dataset 1, Dataset 2, Dataset 1 + Dataset 2) with 3 unsupervised models and obtained independent results w.r.t to the researches done in the domain parameters.

VII. References

[1] Mathur, V. Saxena and S. K. Singh,―Understanding sarcasm in speech using Mel- cepstral frequency coefficient, 2017 7 th International Conference on Cloud Computing, Data Science & Engineering Confluence, Noida, (2017) 728-732. doi: 10.1109/CONFLUENCE.2017.7943246, 2017.

[2] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., and Poria, S., (2019). Towards Multimodal Sarcasm Detection (An_Obviously_ Perfect Paper). [online] arXiv.org. Available at: <https://arxiv.org/abs/1906.01815> [Accessed 20 May 2020].

[3] Swami, S., Khandelwal, A., Singh, V., Akhtar, S., and Shrivastava, M., (2018). A Corpus of English-Hindi Code- Mixed Tweets for Sarcasm Detection. [online] arXiv.org. Available at:

<https://arxiv.org/abs/1805.11869v1> [Accessed 20 May 2020].

[4] Rakov, R. & Rosenberg, A., Sure, I did the right thing": A system for sarcasm detection in speech. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH., (2013) 842-846.

[5] X, L. and X, V., (2019). Sarcasm Detection. [online] Web.stanford.edu. Available at: <https://web.stanford.edu/class/cs224n/posters/1579178 1.pdf> [Accessed 9 April 2021].

[6] J. Jody, How I Understood: What features to consider while training audio files?, Medium, (2021). [Online]. Available: https://towardsdatascience.com/how-i-understood-what- features-to-consider-while-training-audio-files- eedfb6e9002b. [Accessed: 28-Apr- 2021].