

NATURAL LANGUAGE PROCESSING

Arvind Chandrasekaran

Graduate Student, Colorado Technical University, Colorado, USA

ABSTRACT

Hints of the outbreak are detected through the modified circumstances favoring the outbreaks, like the warm weather contributing to epidermal outbreaks or the loss of sanitation leading to cholera outbreaks typically relying on the routine reports from the healthcare facilities, secondary data like attendance monitoring at workplaces and schools, the web, and the media play a significant informational source with more than 60% of the initial outbreak reporting to the informal sources. Through the application of natural language processing methods and machine learning technologies, a pipeline is developed which extracts the critical entities like country, confirmed case counts, disease, and case dates, which are mandatory entities from the epidemiological article and are saved in the database thereby facilitating the data entry easier. The advantages are the facilitation of relevant score articles shown first, thereby providing the web service results termed EventEpi integrated into the Event Based Surveillance (EBS) workflows.

KEYWORDS

Event-based surveillance; Contribution; Methods; Information extraction; Entity filtering; Scoring; Evaluation.

1. INTRODUCTION

The public health surveillance goal is the timely detection and containment of disease outbreaks effectively to minimize the health consequences and the public health burden. The data acquisition under the traditional report systems is a passive process, and the routine is followed and established by the institutes of public health and the legislator. (Abbood; Busche; Ghazzi. et al., Nov 2020) The process is termed indicator-based surveillance. The outbreak hints could be detected through changed circumstances known to favor the outbreaks, for instance, the warm weather contributing to more epidemic outbreaks or the loss of sanitation leading to the cholera outbreak. (Abbood; Busche; Ghazzi. et al., Nov 2020) Massive data filtering needs help finding the proper criteria for the information considered and discarded. The task could get complex as the filter mustn't miss the important events while confidence exists to exclude. (Abbood; Busche; Ghazzi. et al., Nov 2020) Without the above filters, performing EBS for extensive data is not feasible. Natural language processing algorithms are suited well to capture informal resources and guide information filtering and structuring systematically and automatically.

2. SIGN LANGUAGE

Sign Language facilitates the people for the creation of an inclusive society of people having disabilities with equal chances for development and growth to live safe, productive, and dignified life. Sign language isn't included in the teaching materials, and the hard-to-hear children's parents are unaware of the value of sign language for bridging the communication gaps. (Sharma; Tulsian; Verma. et al., 2022) The responsibility of Indian Sign Language encourages hard-of-hearing students in intermediate, primary, and higher education to use Indian Sign Language as an instruction form. They train and educate diverse groups like teachers, government officials, David C. Wyld et al. (Eds): DBML, CITE, IOTBC, EDUPAN, NLP AI -2023 pp. 57-61, 2023. IJCI – 2023

professionals, the public, and community teachers. (Sharma; Tulsian; Verma. et al., 2022) They promote Sign Language in collaboration with hard-of-hearing groups and institutions with disabilities. HamNoSys is the sign language is the International Phonetic Alphabet intended to transcribe the sounds consistently for any spoken language. The parameter versions preclude the well-known alphabet with new glyphs developed to deduce the symbol’s meaning. HamNoSys consists of gestures, pointers, and handshapes. Sign language possesses hand gestures for expressing the words or sentences expressing the implications arising from the sign language in-depth and described under three broad categories one-handed, nonmanual, and two-handed.

3. PROPOSED SYSTEM

The system is proposed using natural language processing capabilities of recognizing, interpreting, and expressing dynamic hand gestures per the ASL standards for ten languages. (Deshpande; Halgekar; Kulkarni. et al., 2021) The system provides the capability to adjust the dynamic gestures for the letter’s proper recognition, thereby preventing the chances of letter recognition as different and remaining the same. (Deshpande; Halgekar; Kulkarni. et al., 2021) They can recognize dynamic hand movements, accurately associate with static hand gestures from the sentence for conversion to other languages and detect the sentiment behind them. The conceptual model for the ER diagram is shown below.

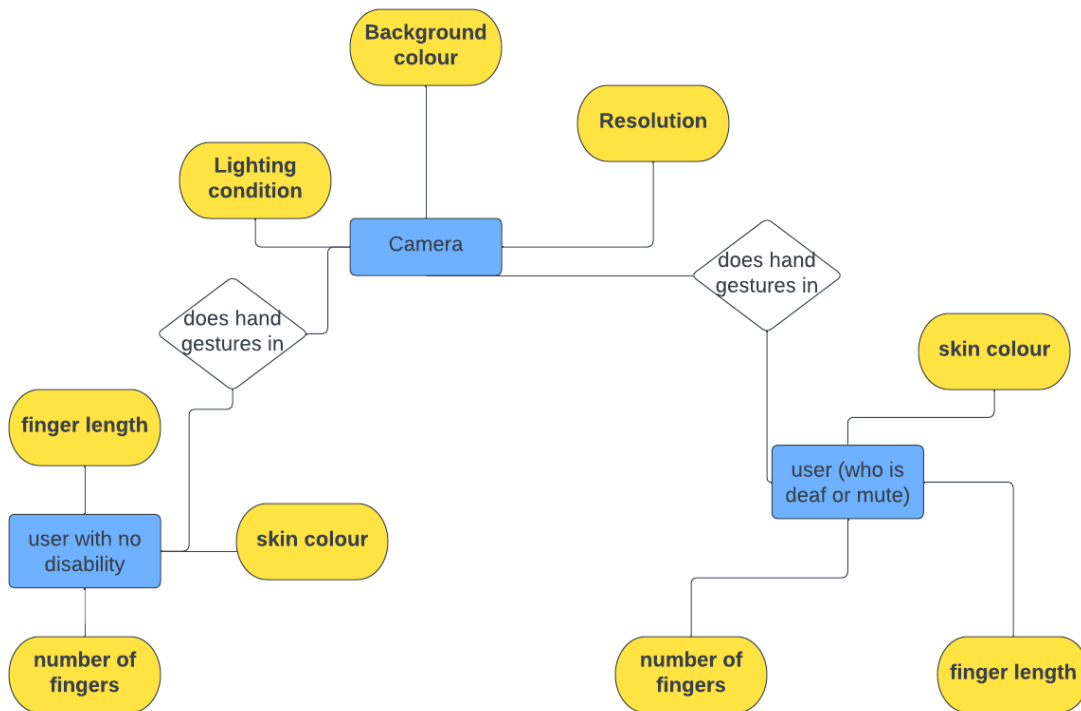


Figure 1. ER Diagram for the system

The above entity relationship diagram is the technique of data modeling graphically illustrating the system’s entities. The system’s entities are the users having disabilities and without disabilities and the webcam laptop.

3.1. Sequence Diagram

The sequence diagram regarding the object interactions arranged as the time sequence is shown below. The diagram depicts the classes and objects involved in the scenario (Deshpande; Halgekar; Kulkarni. et al., 2021), and the message sequences are exchanged between the needed objects to carry out the scenario functionality.

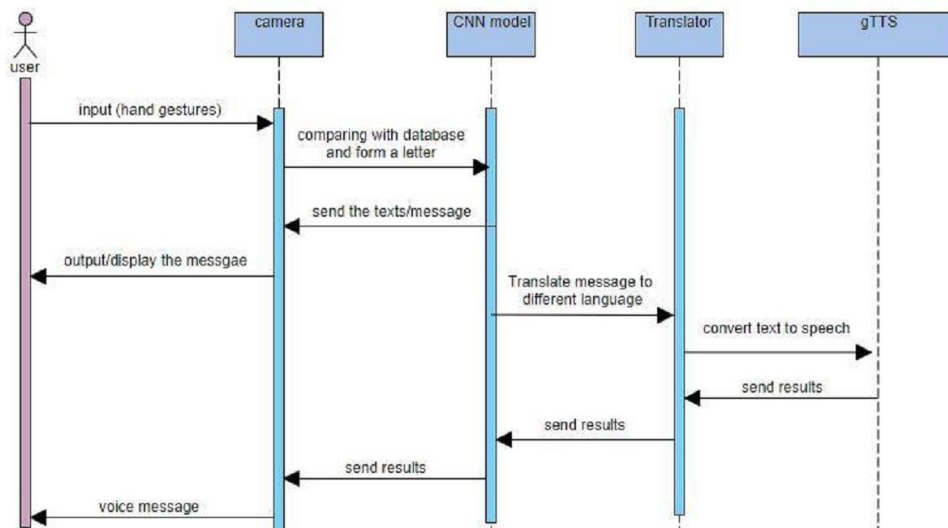


Figure 2. Sequence Diagram for User Module

3.2. Implementation

The convolutional neural network is developed for recognizing ASL gestures through pre-existing American Sign Language datasets. The image processing and image frames are captured through hand gestures for prediction. (Deshpande; Halgekar; Kulkarni. et al., 2021) The algorithm is applied to translate the predicted data into speech. Vader sentiment analysis is applied to the output text to find the sentiment's conveyed through the statement in the percentage format.

4. TEXT ALIGNMENT

Natural Language Processing has the process of determination of minimum text elements matching for two languages in the translated.js set termed text alignment. Extra-linguistic factors like the text data's provenance, topic, and technical structuring influence text alignment. (Manjula; Shivamurthaiah. et al., 2021) If the character encodings are a close translation of each other, the sentences are the smallest AUs text. For the translation of introductory textbooks, the organization comprised of the chapters and parallel texts would be huge proportionally. (Manjula; Shivamurthaiah. et al., 2021) The initial alignment challenge with the development of parallel corpora linked with the related documents.

5. DEEP LEARNING

Through computer technological advancement, natural language processing has evolved into an interdisciplinary research field covering wide areas like human-computer interaction, artificial intelligence, quantitative linguistics, corpus linguistics, and machine translation. (Feng; Shi. et

al., May 2021) NLP modeling research is generally experienced under three phases empiricism, rationalism, and deep learning. Deep learning is a robust NLP development and has evolved as new life in both AI and language studies involving quantitative methods. (Feng; Shi. et al., May 2021) Natural Language Processing deep learning is an interesting and timely publication as they report deep learning and studies about independent areas of NLP, like conversational language understanding, lexical parsing, and analysis.

5.1. Dataset Extraction

Before deep learning model training, the dataset is established following the subject matter. A customized Corpus is generated to fit the research needs. (Andrade-Seagarra; Gabriel A. et al., 2021) Twitter's API and its advantages are used to generate a Corpus after the extraction, analysis, and processing of tweets through the processing, extraction, and tweets analysis through implementation of the script used subsequently for the models training.

5.2. Training

The above-proposed system is developed from the two methodologies of Recurrent Neural Network (RNN) + Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers. (BERT) (Andrade-Seagarra; Gabriel A. et al., 2021) They are the deep learning models with the highest accuracy and the prediction and classification tasks under the sentiment analysis area. Sequence classification is the predictive modeling problem where the input sequence is present as the input sequence over time or space. (Andrade-Seagarra; Gabriel A. et al., 2021) The goal would be to predict the category for the sequence assignment. The classification model starts with the block development allowing the text preprocessing, so that deep cleaning is applied across the original contents. The dataset preprocessing is performed using the stopwords technique to remove the special characters and punctuation symbols irrelevant to the sentence meaning. (Andrade-Seagarra; Gabriel A. et al., 2021) BERT has the library to perform the tokenization and preprocessing tasks termed BERT Tokenizer. The library splits the text into tokens and converts the tokens into tokenized vocabulary indexes. (Andrade-Seagarra; Gabriel A. et al., 2021) They equalize or limit the sentences of equal length, and the attention mask is created where the last parameter is the required component for the model training.

6. CONCLUSION

Two models of NLP deep learning are used for detecting Twitter social network cyberbullying. The models proposed achieves a 5% accuracy improvement. The model RNN+LSTM is the balanced option having an execution time with an accuracy of 91.82% and an execution time of 78 min. RNN+LSTM is not an efficient tool for cyberbullying the classification option because the BERT has a better criterion for detecting harassment. BERT performance is improved by 20% compared with RNN+LSTM. BERT is a pre-trained model, and the analyzed account doesn't contain offensive comments explicitly presenting the challenge for the models to predict harassment.

REFERENCES

- [1] Abbood, Auss; Ullrich, Alexander; Busche, Rüdiger; Ghozzi, Stéphane. PLoS Computational Biology; San Francisco Vol. 16, Iss. 11, (Nov 2020): e1008277. DOI:10.1371/journal.pcbi.1008277. EventEpi —A natural language processing framework for event-based surveillance. <https://www.proquest.com/compscijour/docview/2479465271/4B2F377419C1445DPQ/1?accountid=144789>
- [2] Andrade-Segarra, Diego A; Gabriel A. Le´on-Paredes. International Journal of Advanced Computer Science and Applications; West Yorkshire Vol. 12, Iss. 5, (2021). DOI:10.14569/IJACSA.2021.0120592. Deep Learning-based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks. <https://www.proquest.com/compscijour/docview/2655118531/BBC286109F22434CPQ/9?accountid=144789>
- [3] Feng, Haoda; Shi, Feng. Natural Language Engineering; Cambridge Vol. 27, Iss. 3, (May 2021): 373-375. DOI:10.1017/S1351324919000597. Deep Learning in Natural Language Processing. <https://www.proquest.com/compscijour/docview/2534572957/BBC286109F22434CPQ/6?accountid=144789>
- [4] Kulkarni, Aishwarya; Halgekar, Pranav; Deshpande, Girish R; Rao, Anagha; Dinni, Aishwarya. Turkish Journal of Computer and Mathematics Education; Trabzon Vol. 12, Iss. 10, (2021): 129-137. Dynamic sign language translating system using deep learning and natural language processing. <https://www.proquest.com/compscijour/docview/2623612530/EF7302645A024D66PQ/4?accountid=144789>
- [5] Manjula, S; Shivamurthaiah, M. Turkish Journal of Computer and Mathematics Education; Trabzon Vol. 12, Iss. 13, (2021): 2465-2472. Identification of Languages from The Text Document Using Natural Language Processing System. <https://www.proquest.com/compscijour/docview/2623930195/79824294D4E347C4PQ/5?accountid=144789>
- [6] Sharma, Purushottam; Tulsian, Devesh; Verma, Chaman; Sharma, Pratibha; Nancy, Nancy. Future Internet; Basel Vol. 14, Iss. 9, (2022): 253. DOI:10.3390/fi14090253. Translating Speech to Indian Sign Language Using Natural Language Processing. <https://www.proquest.com/compscijour/docview/2716521607/AA2ACB9B4AB64EF5PQ/2?accountid=144789>

AUTHOR

Arvind Chandrasekaran from Texas, USA. Presently working for PPG Healthcare for the past six years. I have also been pursuing Doctorate in Computer Science (Big Data Analytics) from Colorado Technical University for the past two years, having completed 54 credits; I'm looking to achieve the same by this year.