Report for AI Survey

On

Natural Language Processing (NLP)

BY

Nilanshu Nilay

24901316

M.Tech First Year

Aritificial Intelligence

2024-26

Submitted to

Dr. Satbir Singh

Dr B R Ambedkar National Institute of Technology

Jalandhar,Punjab(144008)

(Center for Artificial Intelligence)

INDEX

OBJECTIVE

Create a report on a field of AI of choice, the current job openings and the skill set required for those jobs.

Introduction to Natural Language Processing

Natural Language Processing (NLP) is a subfield of artificial intelligence that deals with the interaction between computers and humans in natural language. It's a multidisciplinary field that combines computer science, linguistics, and machine learning to enable computers to process, understand, and generate human language. NLP involves various techniques such as text preprocessing, tokenization, sentiment analysis, named entity recognition, and machine translation to extract insights and meaning from text data. By analyzing and interpreting human language, NLP enables applications like language translation, chatbots, sentiment analysis, and voice assistants to understand and respond to human input, revolutionizing the way we interact with technology.

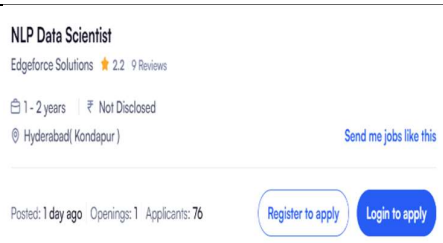
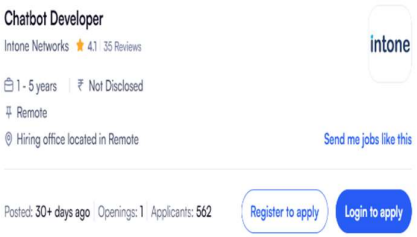
Application of Natural Language Processing

1. **Virtual Assistants**: Apple's Siri, Amazon's Alexa, and Google Assistant use NLP to understand voice commands and respond accordingly.
2. **Language Translation**: Google Translate, Microsoft Translator, and IBM Watson Translation use NLP to translate text and speech across languages.
3. **Sentiment Analysis**: Companies use NLP to analyze customer feedback, reviews, and social media posts to gauge sentiment and improve their products and services.
4. **Text Summarization**: NLP is used to summarize long documents, articles, and websites, highlighting key points and main ideas.
5. **Chatbots**: Chatbots use NLP to understand customer queries and respond with relevant answers, helping with customer support and service.
6. **Speech Recognition**: NLP is used in speech recognition systems to recognize spoken words and phrases, transcribing them into text.
7. **Information Retrieval**: Search engines use NLP to understand search queries and provide relevant results.
8. **Named Entity Recognition**: NLP is used to identify and extract named entities (e.g., people, places, organizations) from text data.
9. **Content Generation**: NLP is used to generate content such as articles, social media posts, and product descriptions.

Additionally, NLP has applications in:

- Healthcare (clinical decision support, medical text analysis)
- Education (intelligent tutoring systems, content analysis)
- Marketing (personalized recommendations, ad targeting)
- Finance (sentiment analysis, risk analysis)

These are just a few examples of the many applications of NLP. The field is constantly evolving, and new applications are emerging as NLP technology improves.

Website	Name of the Organization	Skill set	Location
Naukri.co m		Design, develop, and optimize NLP models for applications like chatbots, virtual assistants, and text analytics. Develop and implement Generative AI models and algorithms using advanced techniques such as GPT, Variational Autoencoders (VAE), and Generative Adversarial Networks (GANs). Conduct research to stay current with the latest advancements in Generative AI, Machine Learning (ML), and Deep Learning (DL) and identify opportunities to integrate them into our products and services.	Hyderabad, India
Naukri.co m		**Design, develop, and deploy AI-powered chatbots for customer service and internal automation.** **Work with business stakeholders to understand the requirements and develop conversation flows and scripts.** **Integrate chatbot solutions with existing platforms, APIs, and databases for seamless data retrieval.** **Implement natural language processing (NLP) algorithms to enable the chatbot to understand and respond to user inputs.**	Remote
Linkedin		**Experience building a LLM or other large models from scratch including writing model code, gathering training data, optimizing training and inference.** **Strong experience in NLP, multimodal learning, proficient in TensorFlow, PyTorch, JAX and CUDA toolkit.**	

Summary of Research Paper

Paper Title:

Live Event Detection for People's Safety Using NLP and Deep Learning

Author and Publication Details:
Authors: AMRIT SEN1, GAYATHRI RAJAKUMARAN 1, MIROSLAV MAHDAL 2, SHOLA USHARANI1,
VEZHAVENDHAN RAJASEKHARAN 3, RAJIV VINCENT 1, AND KARTHIKEYAN SUGAVANAN1

Key points :

1. **Research Objective**: The study aims to develop a software-based system for detecting threats from ambient noise using NLP and deep learning, and automatically alerting registered contacts via email, SMS, and WhatsApp.
2. **Problem Statement**: The system addresses the need for real-time threat detection, especially for individuals working alone at night in remote areas, by analyzing surrounding sounds for potential dangers like robbery, assault, or homicide.
3. **Methodology**:
    - **Data Collection**: Utilized audio signals from Kaggle datasets, consisting of over 9000 audio clips across 13 classes.
    - **Data Analysis**: Applied Exploratory Data Analytics (EDA) techniques and transformed audio data using Fast Fourier Transform and Mel-Spectrograms.
    - **Deep Learning Models**: Trained three models (1D-CNN, 2D-CNN, and LSTM) on the cleansed dataset, achieving high accuracy in classifying audio events.
4. **System Functionality**:
    - **Live Audio Recording**: Records ambient noise and processes it to detect potential threats.
    - **Prediction Module**: Classifies the sound and sends automatic alerts if a threat is detected.
    - **Alert Mechanism**: Sends alerts via email (with audio attachment), SMS, and WhatsApp to registered contacts and emergency services.
5. **Results**:
    - **Accuracy**: Achieved 96.6% accuracy with the LSTM model, 96.3% with the 2D-CNN, and 95.2% with the 1D-CNN.
    - **Confusion Matrix**: Used to evaluate the performance of the models in predicting the correct audio classes.
6. **Future Scope**:
    - **Dataset Expansion**: Increasing the number of audio classes and dataset size for better model performance.
    - **Feature Extraction**: Exploring additional features like amplitude, phase, and harmonic distortion for improved analysis.
    - **Advanced Models**: Researching more complex RNN models like GRU, Bi-GRU, and HRNN for enhanced results.
    - **Additional Functionalities**: Integrating location sharing and other parameters (e.g., heart rate, temperature) for comprehensive emergency response.

7. **Advantages**:
    - **No Additional Hardware**: The system operates using a smartphone's default microphone, eliminating the need for extra devices.
    - **Immediate Alerts**: Provides quick and automatic alerts to ensure timely assistance.

Challenges in the research:

1. **Hardware Dependency**: Previous systems relied on bulky hardware, which posed difficulties in regular use. The current research aimed to eliminate the need for additional hardware by using only a smartphone.
2. **Data Variability**: The dataset consisted of audio files with varying formats, channels, and sampling rates, which required extensive pre-processing to ensure uniform analysis.
3. **Silent Zones in Audio**: Many audio files contained significant silent areas, necessitating the creation of a custom signal envelope to remove dead spaces and focus on relevant audio segments.
4. **Computational Complexity**: Training deep learning models on large datasets with multiple classes required significant computational power and time, making it challenging to achieve quick and efficient training.
5. **Overfitting**: Ensuring that the models did not overfit the training data was a challenge, especially with complex models like LSTM. Regularization techniques and dropout layers were used to mitigate this issue.
6. **Real-time Processing**: Implementing real-time audio recording, processing, and prediction required efficient algorithms and optimization to ensure timely threat detection and alert generation.
7. **Model Selection**: Choosing the appropriate deep learning model (1D-CNN, 2D-CNN, or LSTM) for different scenarios and ensuring each model's performance was a challenge, especially as the dataset size and complexity increased.
8. **Alert Accuracy**: Ensuring the accuracy of threat detection and minimizing false positives/negatives was crucial for the system's reliability and effectiveness in real-world scenarios.
9. **Integration of Additional Features**: Incorporating additional functionalities like location sharing and other physical/environmental parameters required further research and technical implementation.
10. **Scalability**: Scaling the system to handle larger datasets and more audio classes while maintaining high accuracy and performance was a significant challenge.

These challenges highlight the complexity of developing a robust and reliable threat detection system using NLP and deep learning, emphasizing the need for continuous research and improvement.

Conclusion: The research successfully developed a threat detection and alert system using deep learning and audio analysis, providing a practical solution to enhance individual safety in real-time.

Paper Title:

NLP-Based Fusion Approach to Robust Image Captioning

Authors: -

Riccardo Ricci, *Graduate Student Member, IEEE*, Farid Melgani, *Fellow, IEEE*, José Marcato Junior , *Member, IEEE*, and Wesley Nunes Gonçalves

Key points:

The research paper "NLP-Based Fusion Approach to Robust Image Captioning" by Riccardo Ricci, Farid Melgani, José Marcato Junior, and Wesley Nunes Gonçalves focuses on enhancing the robustness of remote sensing image captioning through an ensemble approach. Here are the key points:

1. **Problem Statement**: The paper addresses the challenge of robustness in remote sensing image captioning, particularly when data scarcity limits the generalization capability of single captioning models.
2. **Proposed Solution**: The authors propose an ensemble approach to select or generate the most coherent caption from a set of predictions made by different captioning algorithms. This method aims to improve robustness without significantly increasing complexity.
3. **Ensemble Strategies**:
    - **Naïve Selection**: Uses a BERT model to project captions into a semantic space and selects the caption closest to the average embedding.
    - **CLIP-Coherence Selection**: Utilizes the CLIP model to compute image-text coherence scores and selects the caption with the highest score.
    - **VaE Fusion Strategy**: Employs a variational autoencoder (VaE) to distill a single caption from the set of generated captions.
4. **Datasets**: The study uses four datasets for validation: UCM-Captions, SIDNEY-Captions, RSICD, and UAV-Captions.
5. **Experimental Scenarios**:
    - **Standard Evaluation**: Evaluates the ensemble on the same dataset used for training.
    - **Generalization Evaluation**: Tests the ensemble's performance on datasets different from the training set.
    - **Robustness Evaluation**: Assesses the ensemble's resilience to noise and errors in the input captions.
6. **Results**:
    - The ensemble approach generally improves robustness and performance compared to individual models.
    - The VaE fusion strategy is particularly effective in handling noise and errors.
    - The CLIP-coherence selection strategy shows promise, especially with more specialized CLIP models for remote sensing.
7. **Conclusion**: The ensemble strategies enhance the reliability and contextual relevance of image captions in remote sensing applications. Future research could focus on integrating more specialized CLIP models and automated filtering mechanisms to further improve performance.
8. **Computational Overhead**: The ensemble approach incurs a significant computational overhead, making it suitable for scenarios where time is not a constraint.

The paper highlights the potential of ensemble methods to address the limitations of single-model approaches in remote sensing image captioning, providing a scalable and robust solution for real-world applications.

Paper Title:

AI-Assisted Deep NLP-Based Approach for Prediction of Fake News From Social Media Users

Authors:-

Ganesh Gopal Devarajan, Senior Member, IEEE, Senthil Murugan Nagarajan, Member, IEEE, Sardar Irfanullah Amanullah, S. A. Sahaaya Arul Mary, and Ali Kashif Bashir, Senior Member, IEEE

Key points:-

The research paper "AI-Assisted Deep NLP-Based Approach for Prediction of Fake News From Social Media Users" presents a novel AI-assisted model for detecting fake news using deep natural language processing (NLP). Here are the key points:

1. **Objective**: The paper aims to detect fake news and verify the credibility of social media users and publishers using a deep learning model.
2. **Proposed Model**: The model is called NLP-integrated deep CNN Bi-LSTM (N-DCBL) attention network. It consists of four layers: publisher layer, social media networking layer, enabled edge layer, and cloud layer.
3. **Methodology**:
    - **Data Acquisition**: Collecting data from various sources.
    - **Information Retrieval (IR)**: Extracting relevant information.
    - **NLP-Based Data Processing**: Preprocessing data using techniques like segmentation, cleaning, PoS tagging, stop word removal, and word embedding.
    - **Deep Learning-Based Classification**: Using a deep learning model to classify news articles as fake or real based on various features.
4. **Datasets Used**: The model was evaluated using three datasets: Buzzface, FakeNewsNet, and Twitter.
5. **Performance**: The proposed model achieved an average accuracy of 99.72% and an F1 score of 98.33%, outperforming existing methods.
6. **Architecture**: The model includes a deep CNN Bi-LSTM attention network for fake news detection and a multiheaded attention network for verifying user and publisher credibility.
7. **Evaluation Metrics**: The performance was evaluated using accuracy, recall, precision, F1 measure, and loss.
8. **Results**: The N-DCBL model showed superior performance compared to other deep learning models like CNN, ResNet, and Bi-LSTM.
9. **Conclusion**: The proposed model effectively detects fake news and verifies user credibility, preventing the spread of misinformation on social media.
10. **Future Work**: The paper suggests extending the work by adding effective feature selection methods to avoid overfitting issues.

These points summarize the main contributions and findings of the research paper.

Paper Title:

Medical Information Extraction With NLP-Powered QABots: A Real-World Scenario

Authors: -

Claudio Crema , Federico Verde , Pietro Tiraboschi, Camillo Marra, Andrea Arighi, Silvia Fostinelli , Guido Maria Giuffré , Vera Pacoova Dal Maschio, Federica L'Abbate, Federica Solca, Barbara Poletti , Vincenzo Silani , Emanuela Rotondo , Vittoria Borracci, Roberto Vimercati, Valeria Crepaldi, Emanuela Inguscio, Massimo Filippi , Francesca Caso, Alessandra Maria Rosati, Davide Quaranta, Giuliano Binetti, Ilaria Pagnoni, Manuela Morreale, Francesca Burgio, Michelangelo Stanzani-Maserati Sabina Capellari , Matteo Pardini , Nicola Girtler, Federica Piras , Fabrizio Piras, Stefania Lalli, Elena Perdixi, Gemma Lombardi , Sonia Di Tella, Alfredo Costa, Marco Capelli, Cira Fundarò, Marina Manera, Cristina Muscio, Elisa Pellencin , Raffaele

Abstract

This study presents an innovative approach for extracting medical information from clinical documents using Natural Language Processing (NLP) techniques and Question Answering Bots (QABots). The approach involves defining a common Case Report Form (CRF) for multiple centers and developing the NLP Extraction and Management Tool (NEMT), a semi-automated end-to-end pipeline for extracting relevant information from clinical documents and storing it in a centralized database. The study demonstrates the effectiveness of the proposed data collection and fine-tuning strategy for training NEMT's QABot model, achieving high performance metrics (EM score = 78.1%, F1-score = 84.7%, LAcc = 0.834, MRR = 0.810) on an annotated dataset of clinical documents. The study also highlights the limitations of traditional metrics used to evaluate QABot models and proposes exploring newer metrics that consider the semantic content of predicted answers.

Key Points

• A common Case Report Form (CRF) was defined for multiple centers, and the NLP Extraction and Management Tool (NEMT) was developed for extracting information from clinical documents.
• NEMT is a semi-automated end-to-end pipeline that uses a Question Answering Bot (QABot) to extract relevant information and store it in a centralized database.
• The study demonstrated the effectiveness of the proposed data collection and fine-tuning strategy for training NEMT's QABot model.
• The study achieved high performance metrics (EM score = 78.1%, F1-score = 84.7%, LAcc = 0.834, MRR = 0.810) on an annotated dataset of clinical documents.
• The Inter-Annotator Agreement (IAA) was calculated to evaluate the reliability of annotations, with an average IAA score of 56% and a macro F1-score of 79%.
• The study highlights the limitations of traditional metrics used to evaluate QABot models and proposes exploring newer metrics that consider the semantic content of predicted answers.
• The study also emphasizes the importance of considering the variability in annotations and the potential for disagreement among annotators.
• A comparison with Large Language Models (LLMs) showed that the proposed QABot model outperforms the LLMs on specific topics.

Paper Title:

Climate Change Sentiment Analysis Using Domain Specific Bidirectional Encoder Representations From Transformers

Authors: -
V. S. ANOOP 1, T. K. AJAY KRISHNAN1, ALI DAUD 2, AMEEN BANJAR 3, AND AMAL BUKHARI

Abstract

This research explores the use of natural language processing and machine learning techniques to analyze public sentiment towards climate change using Twitter data. The study employs ClimateBERT, a domain-specific pre-trained language model, to generate context vectors and classify tweets into positive, negative, and neutral sentiment categories. By collecting and manually labeling 5,506 tweets between January 2022 and February 2023, the researchers developed a comprehensive approach to understand public discourse on climate change, demonstrating the potential of advanced NLP techniques to extract meaningful insights from social media data.

Key Points

- The research collected and manually labeled 5,506 tweets related to climate change using snscrape and Label Studio, with an inter-annotator agreement of 0.72.

- ClimateBERT, a pre-trained transformer model specifically trained on climate-related documents, was used to generate context vectors for sentiment analysis.

- The Random Forest classifier with ClimateBERT embeddings achieved the highest performance, with 90.22% accuracy, 85.73% precision, 85.22% recall, and 85.47% F-measure.

- The study highlights the importance of understanding public sentiment towards climate change for informing policy decisions and stakeholder engagement.

- The research demonstrates the potential of domain-specific language models in capturing nuanced contextual representations for text classification tasks.

- The authors made their labeled dataset publicly available to encourage further research in climate change discourse analysis.

- The approach preprocessed tweets by removing special characters, stopwords, and applying stemming and lemmatization techniques.

Challenges –

1. Data Imbalance:
    - The dataset collected from Twitter exhibited significant class imbalance, with some sentiment categories (positive, negative, or neutral) overrepresented, potentially biasing the model's predictions.
2. Limited Dataset Size:
    - The study relied on only 5,506 labeled tweets, which, despite manual annotation, may not comprehensively represent the vast scope of public sentiment on climate change.
3. Annotation Consistency:
    - Moderate inter-annotator agreement (Krippendorff's alpha = 0.72) reflects challenges in maintaining consistent labeling of tweets, given the subjective nature of sentiment interpretation.
4. Domain-Specific Limitations:
    - While ClimateBERT performs well on climate-related tasks, its reliance on pre-training specific to climate data may limit its generalization to broader NLP tasks.
5. Feature Encoding Trade-offs:
    - Testing multiple encoding techniques (e.g., TF-IDF, Word2Vec, BERT, and ClimateBERT) revealed varying levels of performance. While ClimateBERT outperformed others, the need to explore advanced encodings remains for further improvement.
6. Computational Resources:
    - Training and fine-tuning the ClimateBERT model required significant computational resources, such as NVIDIA A100 GPUs, potentially limiting accessibility for researchers with less advanced infrastructure.
7. Generalization and Scalability:
    - The study's results are promising but were conducted on a relatively small dataset, raising questions about generalization and scalability to more extensive datasets or other domains.
8. Noise in Social Media Data:
    - Tweets often include slang, emojis, hashtags, and irregular grammar, making preprocessing essential but also challenging. Ensuring noise removal without losing valuable context is difficult.
9. Sentiment Complexity:
    - Climate change discussions are nuanced, often blending factual statements with subtle emotions, making accurate sentiment detection complex.
10. Adaptation for Real-World Applications:
    - The paper highlights the need for future work to integrate findings into real-world systems, such as policymaking or large-scale public opinion monitoring, which remain unexplored.

Future Scope of Climate Change Sentiment Analysis Using Domain-Specific Bidirectional Encoder Representations From Transformers

1. Enhanced Dataset Collection:
    - Expand the dataset to include a broader range of sources, such as news articles, scientific reports, and discussions from other social media platforms like Facebook, Reddit, and YouTube.

    - Incorporate multilingual datasets to analyze climate sentiment globally.
2. Improved Annotation Practices:
    - Employ advanced tools like active learning to reduce manual effort and improve annotation quality.
    - Introduce collaborative annotation techniques to achieve higher inter-annotator agreement.
3. Advanced Models and Techniques:
    - Explore deep learning architectures such as transformers with sequential modeling (e.g., RNNs, LSTMs) or hybrid models for improved sentiment detection.
    - Fine-tune domain-specific models like ClimateBERT further with larger and more diverse datasets for better contextual understanding.
4. Real-Time Sentiment Analysis:
    - Develop a framework for real-time monitoring of climate-related sentiment to aid policymakers, NGOs, and researchers in timely decision-making.
    - Integrate sentiment analysis with event detection to identify and analyze spikes in public discourse during climate-related events (e.g., natural disasters or climate summits).
5. Cross-Domain Applicability:
    - Extend the methodology to other domains such as public health, education, and politics to analyze sentiment trends on critical global issues.
6. Multi-Modality Analysis:
    - Combine textual data with images, videos, and geospatial data to better understand sentiment and context in climate discussions.
    - Leverage multi-modal approaches to capture a more comprehensive view of public opinion.
7. Integration with Policy and Communication:
    - Utilize sentiment analysis to tailor climate communication strategies for better public engagement and awareness.
    - Support policymakers with data-driven insights into public concerns, perceptions, and stakeholder influence.

Conclusion-

This study successfully demonstrates the efficacy of ClimateBERT, a domain-specific pre-trained language model, in accurately classifying public sentiment on climate change through the analysis of Twitter data. By integrating ClimateBERT embeddings with machine learning classifiers, particularly the Random Forest model, the research achieved a notable accuracy of 90.22% and an F1-score of 85.47%, outperforming traditional encoding techniques and general-purpose models. The findings provide valuable insights into public perceptions and sentiments regarding climate change, which can inform policymakers and organizations in crafting targeted communication and intervention strategies. Despite challenges such as data imbalance and limited dataset size, the research lays a strong foundation for future advancements in sentiment analysis within the climate change domain. Additionally, by making the labeled dataset publicly available, the study encourages further exploration and enhancement of NLP techniques, ultimately contributing to more informed and effective responses to the global climate crisis.

References of Research Paper

1. https://ieeexplore.ieee.org/ielx7/6287639/10380310/10379088.pdf?tp=&arnumber=10379088&isnumber=10380310&ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzEwMzc5MDg4
2. https://ieeexplore.ieee.org/ielx8/4609443/10330207/10555151.pdf?tp=&arnumber=10555151&isnumber=10330207&ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzEwNTU1MTUx
3. https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10086954&ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzEwMDg2OTU0
4. https://ieeexplore.ieee.org/ielx8/6221020/10745910/10651607.pdf?tp=&arnumber=10651607&isnumber=10745910&ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzEwNjUxNjA3&tag=1
5. https://ieeexplore.ieee.org/ielx8/6287639/10380310/10632142.pdf?tp=&arnumber=10632142&isnumber=10380310&ref=aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzEwNjMyMTQy&tag=1