Name: Aaditya Pal

Rollno:**4045A029**

Class: TYBSCIT

Control id:2021080314

Topic: Natural Language Processing (NLP)

ABSTRACT:

Natural Language Processing (NLP) is a sub-branch of artificial intelligence that is concerned with computers being able to understand, process, and generate human languages. It consists of various types of language models that have evolved with time and have become increasingly advanced to the point where tools such as ChatGPT or various chatbots can engage in a conversation with a human and the human won't be able to tell the difference whether they are communicating with a human or a machine.

INTRODUCTION:

NLP is at the center of AI, as humans interact with it the most. It is also very accessible to everyone. NLP has become the talk of the town since pre-trained transformer models such as GPT-3 were introduced which revolutionized NLP and its applications which are now driven by transformer-based agents.
Any language is governed by a set of rules and symbols. The primary objective of NLP's existence is to bridge the gap between the human language and the computer understandable language. It aims to ease the process of communicating with the computer and provide a more natural way for a human to be able to communicate with a machine. NLP also includes linguistics which is the science of language [1]. NLP can be broadly broken down into two parts- Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU deals with understanding the text in natural languages along with their context and the meaning that is indirectly implied (read between the lines) by them. NLG deals with the generation of text in a natural language. The hue introduction of GPT-3 was the game changer in NLP and brought transformer-based models into the limelight. The success of GPT-3, saw it being adopted in multiple applications. It improved the existing applications of NLP while also giving rise to new applications since accessing NLP became easier. But as the number of applications grew, it gave rise to its own set of new challenges along the way. Every coin has two sides, and understanding its challenges and negatives is equally important rather than just focusing on its benefits.

DETAILED REVIEW:

Review 1

What is NLP?

Diksha Khurana et al. [1] have classified Natural Language Processing(NLP) into 2 parts namely Natural Language Understanding and Natural Language Generation.

- Natural Language Understanding :
  It refers to machines being able to understand and analyze natural language by extracting its underlying context and meaning.
  There are some terms used in different levels of NLP.

- Phonology:- It refers to the systematic arrangement of sound and is considered as a subpart of linguistics[1].

- Morphology:- A word can be broken down into various parts and each part can convey some meaning. These parts are called morphemes. The words that can't be broken down are called Lexical morphemes. The words that are combined with the lexical morpheme are known as Grammatical morpheme [1]. Grammatical morphemes that need to be bound with other words in order to appear as an entire word are known as Bound morphemes.
  Bound morphemes are further divided into inflectional and derivational morphemes. Inflectional morphemes change the tense, gender, person, etc. when added to a word. Derivational morphemes change the meaning of the word when used. [1]

- Lexical:-
  It refers to systems and humans interpreting the meaning of individual words. A part-of-speech tag is assigned to each word. If the word can behave as multiple parts of speech -e.g. it can be a verb and a noun, the most likely part of speech is taken into consideration that occurs in the particular context. This tag is used for cleaning by removing stop words, stemming, and lemmatization. Stop words are words such as 'in', 'the', 'and' [1] which can be removed as they don't mean anything as single words and appear multiple times. Stemming refers to removing suffixes from the word in order to get the base word. Lemmatization refers to getting the base word not by removing suffixes but instead by getting the correct base word.

- Syntactic
  It refers to the formation of sentences which is grammatically correct. It takes into account the part of speech, stop words, etc. Stemming and lemmatization can't be applied here as they might change the meaning of the sentence.

- Semantic
  It refers to determining the correct meaning of the sentence. For this purpose, the systems need to process the structure of the sentence and need to look out for the words that are usually used in that context. For eg: when the sentence is about a farmer, the terms fields, crops, pesticides, fertilizers, etc. come along with it. A sentence such as "A butterfly climbs a tree" is syntactically correct but semantically incorrect.

- Discourse
  It aims to understand the relations between two or more sentences. For this purpose it uncovers the linguistic structure at two levels: Anaphora Resolution and Coreference Resolution. Anaphora resolution refers to recognizing the entity referenced with the same anaphor to resolve the references within the sentences. E.g.: i. Mango is sweet. ii. It is yellow in color. Here the word "it" refers to the "mango" and this needs to be understood by the system. Coreference resolution refers to finding all expressions that refer to the same entity.

- Pragmatic
  It's not always necessary that all the knowledge required to understand the sentence is available in the given text, rather some of the information comes from outside. This level is

focused on real-world knowledge. For example, “Do you know what time it is” [1], semantically means that the “current time” is being asked. But in a different context, it can inferred that someone is excited just before something special might happen.

- Natural Language Generation:
  Natural Language Generation refers to the formation of natural language sentences that can be understood by entities that can process natural language. These entities could be humans or other NLP systems. It occurs in four stages: identifying goals, planning on how to achieve these goals, realizing the available communicative resources, and executing the plan to produce the text [1].
  - Speaker and Generator
    An application is required that can generate the required text and a speaker that takes part in text generation.
  - Component and levels of representation
    It comprises of four tasks:
    - Content selection: The information that needs to be included in the text.
    - Textual organization: It refers to how the text is arranged in a logical manner, following grammar.
  - Application or speaker
    The speaker initiates the generation and doesn’t take part in the generation.

Review 2

ChatGPT, Transformers, and pre-trained language models

ChatGPT stands for Generative Pre-trained Transformer and is a generative AI model based on NLP techniques which was developed by OpenAI [2]. It revolutionized the way NLP tasks were carried out and it was trained on a massive dataset comprised of knowledge from the Internet. In May 2020, OpenAI released GPT-3 which was then followed by GPT-4 which was released very recently in March 2023. As compared to GPT-3 which has over 175 billion parameters, GPT-4 raised that bar and has over 100 trillion parameters which brings great improvements in terms of accepting image inputs as well. GPT-4 has already outperformed previous language models and did excellent on the traditional NLP benchmarks. GPT-4 can adopt self-supervising learning with a 30% performance boost using the “Reflexion” technique which produces even more accurate results [2]. This enables AI agents to possess human qualities of self-reflection and self-evaluation. Additionally, GPT-4 can create tests to analyze its own answers and revise its solutions accordingly.

ChatGPT is based on the Transformer architecture which is very commonly used as the baseline structure. It overcomes the limitations of traditional models like Recurrent neural networks (RNN) in managing contextual information. The Transformer consists of two parts: an encoder and a decoder. The encode encodes the provided inputs into hidden representations. These representations are then decoded by the decoder to produce the output. Each layer of the encoder and decoder contains a multi-head attention mechanism and a feed-forward neural network [2]. This multi-head attention mechanism helps to decide which parts of the input sequence are the most important. This helps the model to focus on what’s important and understand complex sentences better. The Transformer can also carry out multiple calculations at once which enables it to handle a lot of data. The Transformer architecture is divided into two categories based on their

training tasks: autoregressive language modeling and masked language modeling [2].
Masked Language Modeling: examples are BERT and RoBERTa which learn by guessing missing words in a sentence by looking at the words around the blank spaces.

Autoregressive language modeling: Implemented by GPT language model variants, the next word in the sequence is predicted based on the words that came before it, following a left-to-right approach. Hence, the generated text sounds more natural when a left-to-right approach is followed.

Transformer Architecture [2]

Review 3

GPT-3

GPT-3 stands for Generative Pre-trained Transformer 3(GPT-3) and is a third-generation, autoregressive language model developed by OpenAI [3]. It is considered as one of the largest and most powerful language model with over 175 billion parameters. It is widely used in NLP for the translation of text, summarization, grammar correction, question answering, and generating original text such as emails, songs, etc. by using deep learning [3]. It takes input which is also called a prompt. The model is extensively trained on a very large textual dataset in English and other languages and is the successor to GPT-2. The most defining feature of this language model is the generation of human-like text. Given the input, it can produce excellent-quality texts which are indistinguishable from an experienced writer. The existing human writers can leverage this power of GPT-3 to level up their content quality. Given its exceptional features, it makes the cut for a valuable tool that can be used for content creation, customer service, and educational purposes.

With these impressive features, GPT-3 comes with a number of limitations. The major challenge is that GPT-3 is trained on information available on the internet. This means that it learns from the internet, which includes people's opinions and societal bias. The internet doesn't always contain facts and these comments can prove quite offensive which may include discrimination based on race, gender, religion, etc. Luciano Floridi et al.[3] ran 3 tests to test the mathematical, semantical, and ethical capabilities of GPT-3 and found the above to be true. For example, if GPT-3 is trained using sexist language, GPT-3 may generate text that is sexist or reinforces gender stereotypes. The generation of fake news will also become very easy

Another concern is that it can replace the jobs done by humans. For eg, Microsoft fired a majority of journalists and replaced them with automatic systems for the production of news on MSN[3]. The writers need to learn about prompts and get good at it, in order to keep their jobs.

This quality of content is indistinguishable from human-generated content and there might be revision in the rules for Nobel Prize and other prestigious awards that need to be careful while choosing the right candidate. As the content generation will become easier, the physical memory to store it will come under pressure.

The software engineering jobs will also come under threat, as GPT-3 can also write software code. This will enable marketers to create applications to automate various tasks. So there are two sides of the same coin.

With all these challenges and benefits of GPT-3, Luciano Floridi et al.[3] say that humanity needs to act responsibly and critically. Better copyright legislation and mechanisms to determine the authenticity of the articles and other important documents would be required. The benefits of GPT-3 should be used for the benefit of society and not for other unethical purposes. Ongoing research and development are needed to improve GPT-3's understanding of language and context and to ensure that it generates text that is fair and follows the laws in place related to discrimination and other human rights. [3]

Review 4

Applications of NLP

Natural Language Processing (NLP) is a powerful tool that is used widely for many purposes such as:

1. Machine Translation
   It refers to translating sentences from one natural language to another natural language while maintaining the meaning and context of the sentences. This translation is done by machines or engines such as Google Translate. Google, the software announced that it would be using deep learning advanced NLP and artificial neural networks to introduce a new machine translation system. [1]

2. Text Categorization and Analysis
   Massive amounts of information are available in the form of texts and going through them manually takes up a lot of time. NLP speeds up this process by categorizing sentences based on the filters set on the provided data set, tagging the sentences, highlighting the sentences, etc. A very good example would be plagiarism checkers such as Turnitin, which check for plagiarism and accordingly highlight the sentences that may be plagiarised.

3. Spam Filtering
   Spam filtering is widely used to filter out spam emails. The system goes through the text and uses algorithms and protocols to determine whether the given email is spam or not. There are various types of spam filters available such as Content filters, Header filters, General Blacklist filters, Rule Based Filters, Permission Filters, and Challenge Response Filters. [1]

4. Information Extraction and summarization
   A lot of useful knowledge and information is stored in the form of text. Going through massive data stores of text is very time-consuming and tedious. Information extraction is a great way of collecting data as well. For eg: to know about the various departments of the Indian Government and their roles, one can mine the data from the official website and process it using NLP. Also, information extraction is used by search engines to extract recent searches and then display ads based on them. Another example is of MITA (Metlife's Intelligent Text Analyzer) [1], which extracts information from life insurance applications. This extracted data can then be processed and summarized in order to understand it.

5. Dialogue System and chatbots
   Chatbots are the most prominent use of NLP. On the Indian Railways website, there is a bot called Disha that guides users on how to book tickets and resolves their queries. Chatbots are a very effective way to provide customer service as well. The recorded calls can be analyzed to train the model to understand the feelings of the customer as well when communicating through calls. The dialogue systems are fully automated systems that can interact with humans in natural languages such as Google Assistant, Microsoft Cortana, Amazon Alexa, and Apple's Siri.

6. Education
   NLP can be used by educational institutes to evaluate online exams and assess the performance of the student. Nowadays, Massive Open Online Courses (MOOC) [4], have emerged which provide various online courses. NLP is used here to extract and understand the customer reviews of the courses and use this extracted information to design better courses and recommend the books related to the courses.

Review 5

Challenges of NLP

Zakaria Kaddari et al. [5] stated the challenges related to Natural Language Processing(NLP). They are as follows:-

1. Challenges related to Natural Language Understanding(NLU):
   NLU comes under NLP, which is concerned with understanding human language and extracting the meaning of the sentence.
   The challenges that occur are:-

   a. Extracting the meaning of a word according to its context.

   b. Understanding the context.

   c. Extracting the semantic meaning and understanding the relationships between synonyms, antonyms, etc

   d. Multi-hop reasoning [5] is related to question-answers and is the ability to answer the question by gathering information from different parts of the text.

   e. Understanding emotions behind the sentences.

   f. Having the knowledge of the world which is also termed as common sense.

   g. Understanding the various ways in which the same sentence can be said.

2. Challenges related to Natural Language Generation(NLG):
   NLG is the task of generating sentences that a human can understand. The transformer-based languages like GPT have become quite advanced in generating natural languages but still, the accuracy and consistency of the generated texts is questioned due to a shortage of data sets.

3. NLP for Low- resource languages
   The rare languages which are not spoken widely around the world are known as low-resource languages. There isn't enough data to train the models for them to understand and process these languages.

4. Evaluation standards
   The NLP models with time have learned to exploit the statistical patterns instead of learning. Hence stricter and better benchmarks are required to find these limitations.

5. Explaining NLP
   Zakaria Kaddari et al.[5] state that there isn't much explanation given for the predictions made by the NLP models and techniques such as sensitive analysis are used to overcome the issue.

6. Multimodal Understanding [5]
   Humans use their eyes, ears, and other senses to determine the context of a particular conversation. On the other hand, the systems only process the text input they are given. This is a challenge as it causes inaccuracy in determining the context.

CONCLUSION:

In conclusion, NLP is at the very center of AI and humans interact with it extensively on a daily basis. NLP standing for Natural Language Processing, is mainly divided into two sections namely Natural Language Generation and Natural Language Understanding. NLU deals with understanding the language whereas the NLG deals with the generation of natural language and both combined with the understanding of linguistic knowledge in order to generate accurate results. With the introduction of GPT-3, this accuracy is taken to another level with Transformer Models taking the limelight. The encoder, decoder, and multi-head attention mechanism form the baseline of these models. These models can be categorized into two types based on the way they are trained. They are Autoregressive language modeling and Masked Language Modeling. GPT-3 is based on the Autoregressive Language Model and it has changed the NLP space by not only understanding the natural language far better than previous models but also generating text that is indistinguishable from human-generated content. This exceptional ability has led to its widespread popularity and can be used to generate content, automate customer service, etc. Due to this outburst of AI, NLP is and can be applied to a number of fields in order to automate the processes. It can be used for extracting information from large texts, summarizing it, understanding customer reviews, improving chatbots and dialogue systems, in the healthcare sector, the education sector, and much more. With all the benefits, there are also the challenges that come along with them. The human language is quite advanced and filled with hidden meanings while coming in different flavors. For a system, to not only understand the words but to understand the meaning of the words in that context is a huge challenge. For a system to be capable of understanding the complex natural language, it needs to be trained. But there is a lack of resources for the rarer languages, hence the models can't be trained in those languages. Also, understanding the emotion behind the words plays quite an important role and is also a major challenge for NLP. All-in-all, NLP is a very powerful tool and can revolutionize how things currently operate and should be used responsibly.

FUTURE SCOPE:

With the outburst of AI, NLP will be used to automate tasks like customer helplines, answering doubts of customers on online websites clearing doubts of students in the education field, and much more. Content generation has seen a steep rise and to speed up the process NLP will play a major role. Support for multiple languages would be available which would help to reach a massive audience. The translation tools would become more sophisticated and that would help remove the language barrier among humans and systems.

REFERENCES:

[1] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh: Natural language processing: state of the art, current trends and challenges
https://doi.org/10.1007/s11042-022-13428-4

[2] Walid Hariri: Unlocking The Potential Of Chatgpt: A Comprehensive Exploration Of Its Applications, Advantages, Limitations, And Future Directions In Natural Language Processing
https://doi.org/10.48550/arXiv.2304.02017

[3] Luciano Floridi, Massimo Chiriatti: GPT-3: Its Nature, Scope, Limits, and Consequences

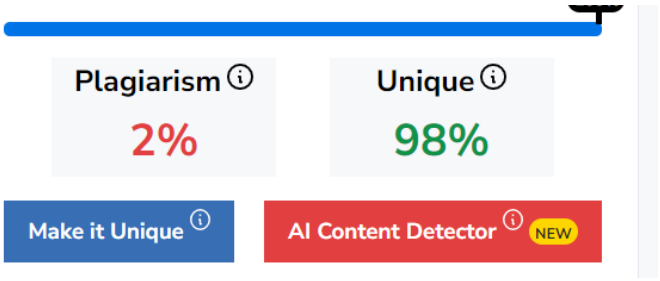
https://doi.org/10.1007/s11023-020-09548-1

[4] Ghania Khensous, Kaouter Labed, Zohra Labed: Exploring the evolution and applications of natural language processing in education
https://rria.ici.ro/wp-content/uploads/2023/06/art._Khensous_Labed_Labed.pdf

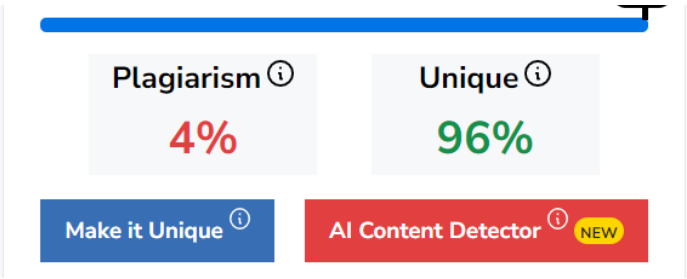
[5] Zakaria Kaddari, Youssef Mellah, Jamal Berrich, Mohammed G. Belkasmi, and Toumi Bouchentouf : Natural Language Processing: Challenges and Future Directions
https://doi.org/10.1007/978-3-030-53970-2_22

PLAGIARISM REPORT:

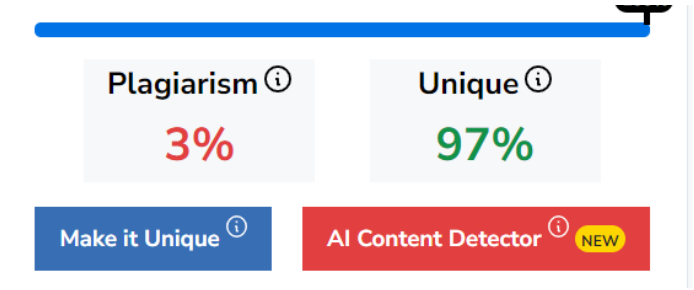
Review 1



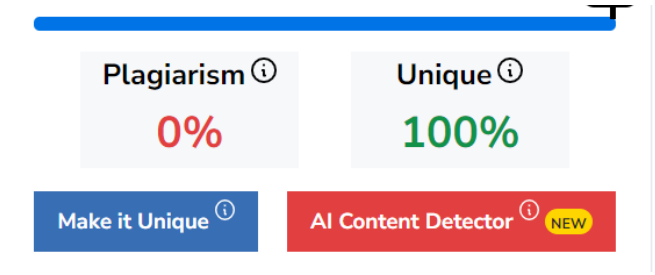
Review 2



Review 3



Review 4



Review 5