UNISA School of Engineering
CSET
FLORIDA / JOHANNESBURG

2023

Prof. M. Sumbwanyambe (SOE)

Open Rubric

NATURAL LANGUAGE PROCESSING (NLP) FOR THE DECOLONISATION OF LOW RESOURCED AFRICAN LANGUAGES

TABLE OF CONTENTS

1. INTRODUCTION TO AI AND ITS APPLICATIONS
2. AI AND AFRICAN CONTEXTUALIZATION. THE MILESTONES.
3. COLONISATION OF AFRICAN LANGUAGES
4. CURRENT RESEARCH IN AFRICAN CONTEXT
5. BENEFITS OF DECOLONISATION OF AFRICAN LANGUAGES.

INTRODUCTION TO AI

What is Artificial intelligence (AI)?

The study of computer systems that attempt to model and apply the intelligence of the human mind

For example, writing a program to pick out objects in a picture

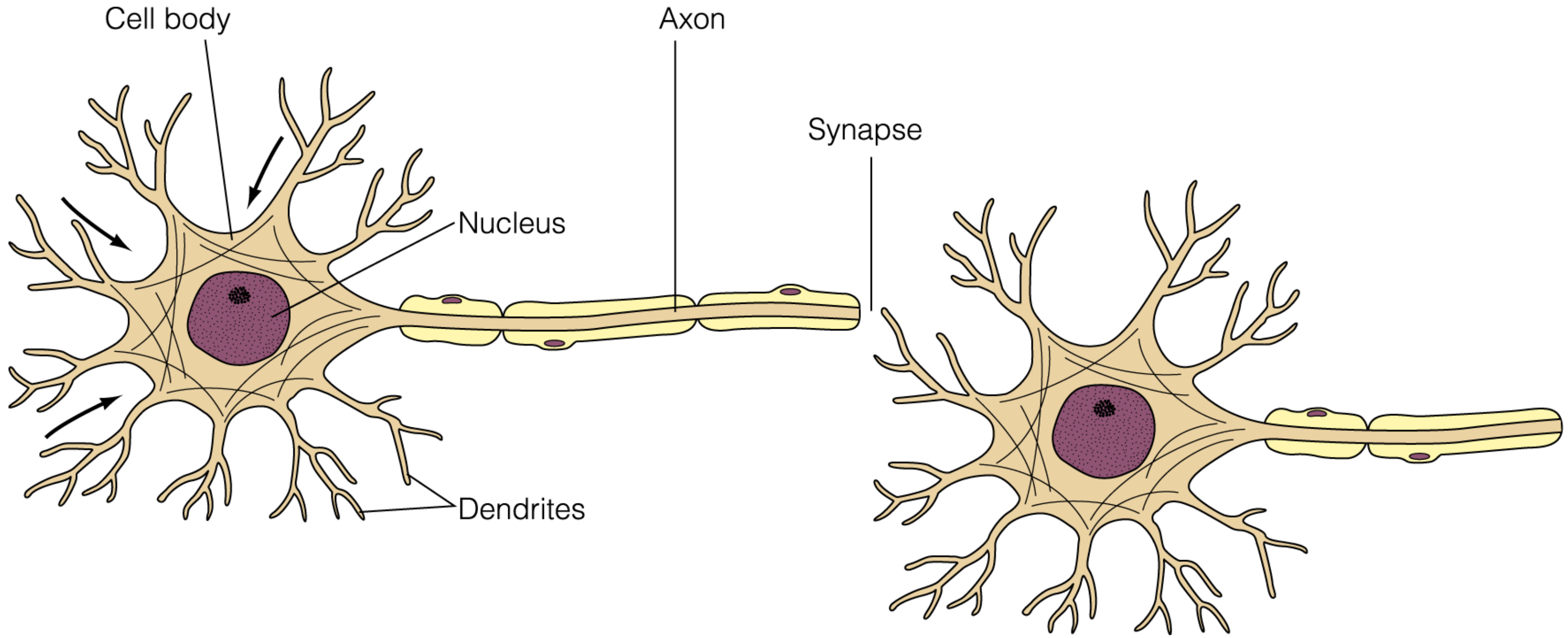
It relates to the bio-Neural network found in a human body.

BRIEF UNDERSTANDING OF A BIO-NEURAL NETWORK

- A series of connected neurons forms a pathway
- A series of excited neurons creates a strong pathway
- A biological neuron has multiple input tentacles called dendrites and one primary output tentacle called an axon
- The gap between an axon and a dendrite is called a synapse

A BIOLOGICAL NEURAL NETWORK.

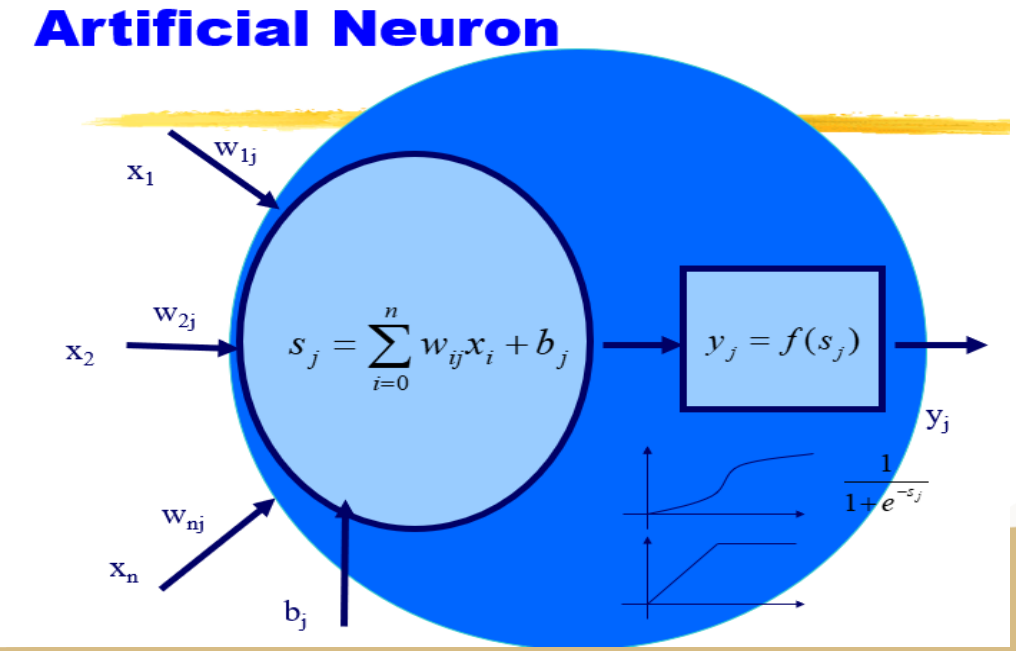
- A biological neural network.



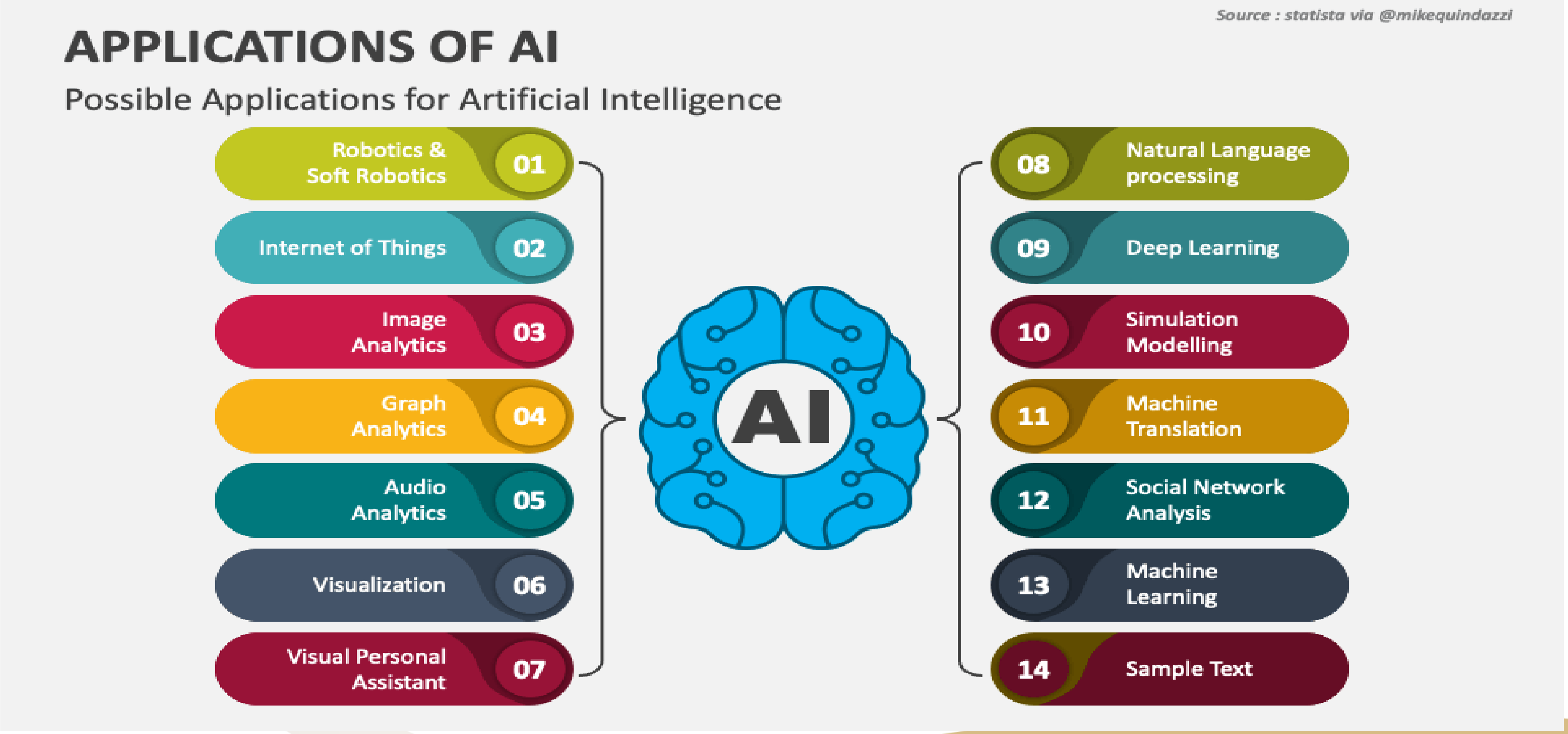
ARTIFICIAL NEURAL NETWORKS

Artificial neural networks

A computer representation of knowledge that attempts to mimic the neural networks of the human body



USES OF AI



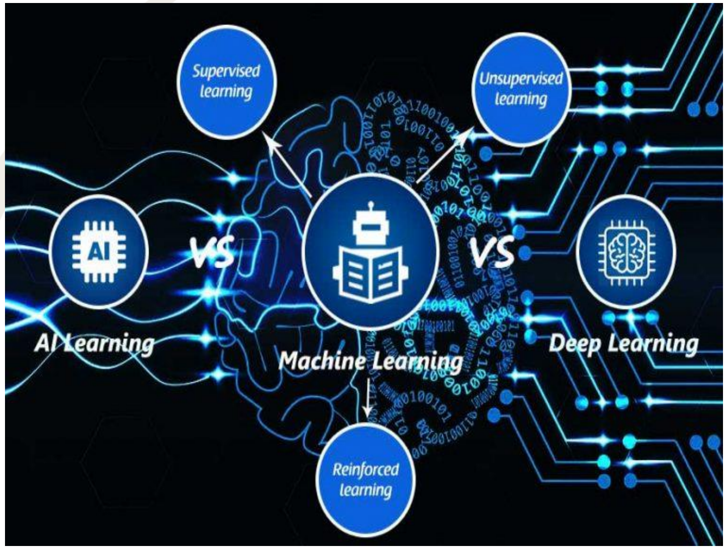
Source : statista via @mikequindazzi

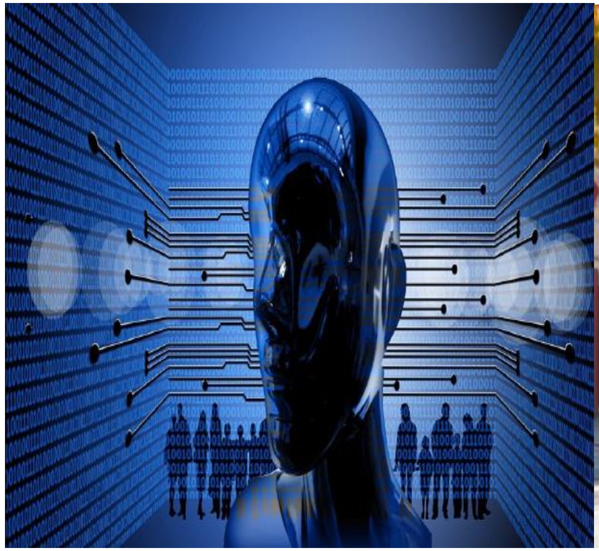
APPLICATIONS OF AI

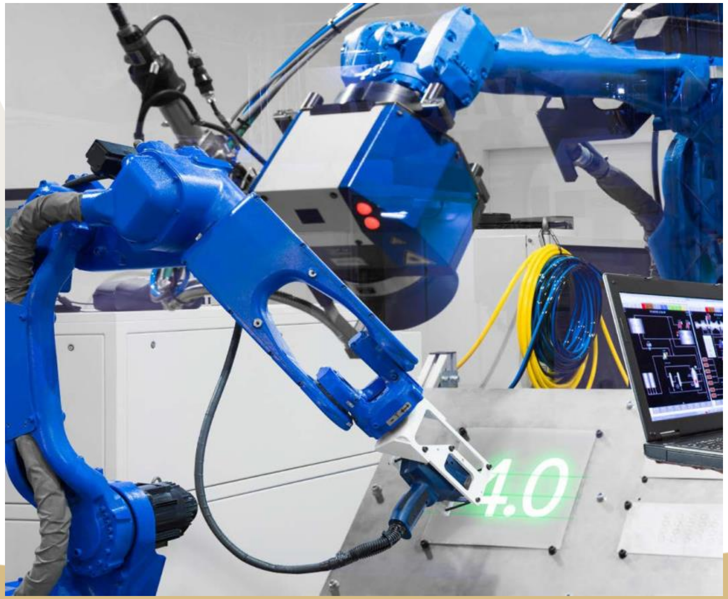
Possible Applications for Artificial Intelligence

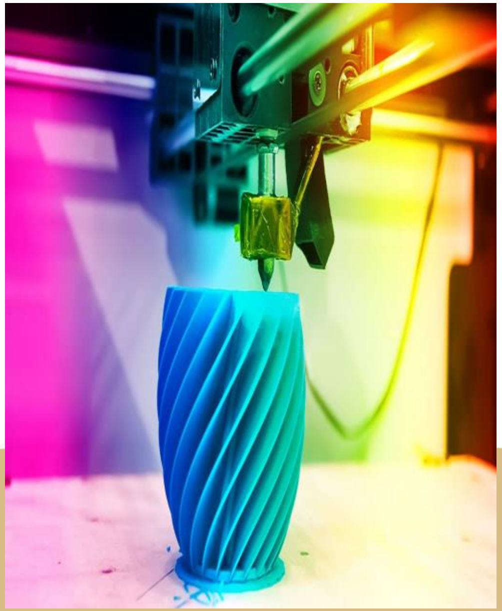
- 01 Robotics & Soft Robotics
- 02 Internet of Things
- 03 Image Analytics
- 04 Graph Analytics
- 05 Audio Analytics
- 06 Visualization
- 07 Visual Personal Assistant
- 08 Natural Language processing
- 09 Deep Learning
- 10 Simulation Modelling
- 11 Machine Translation
- 12 Social Network Analysis
- 13 Machine Learning
- 14 Sample Text

THE FUTURE UNAVOIDABLE: AI APPLICATIONS.



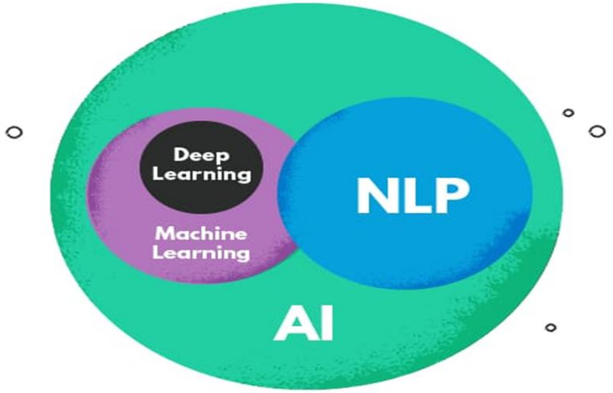






Artificial intelligence a broader perspective

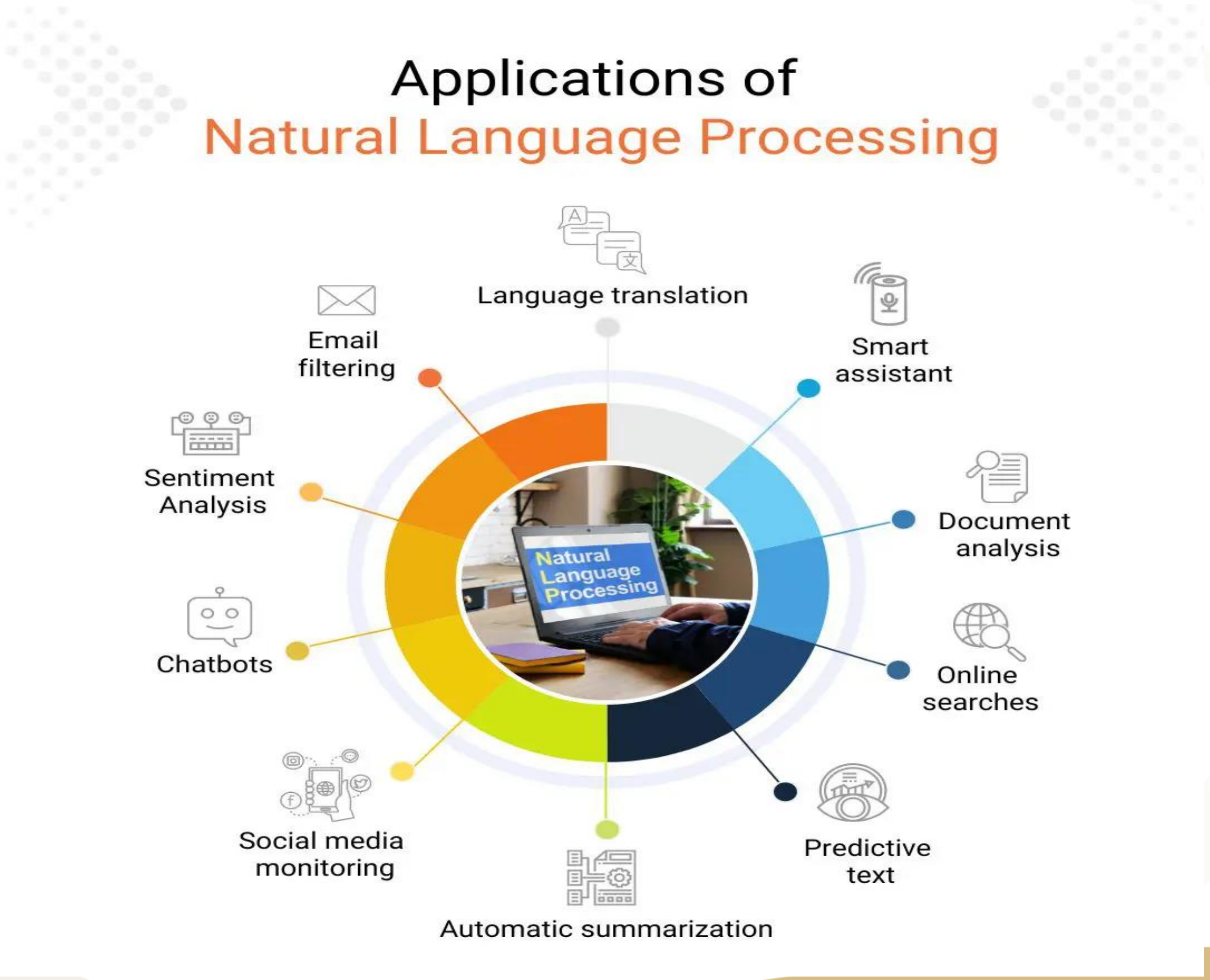
- **Artificial intelligence** (**AI**) is intelligence demonstrated by machines, as opposed to intelligence displayed by humans or by other animals. "Intelligence" encompasses the ability to learn and to reason, to generalize, and to infer meaning**.** Machine learning, deep learning and natural language processing (NLP) are all part of AI.



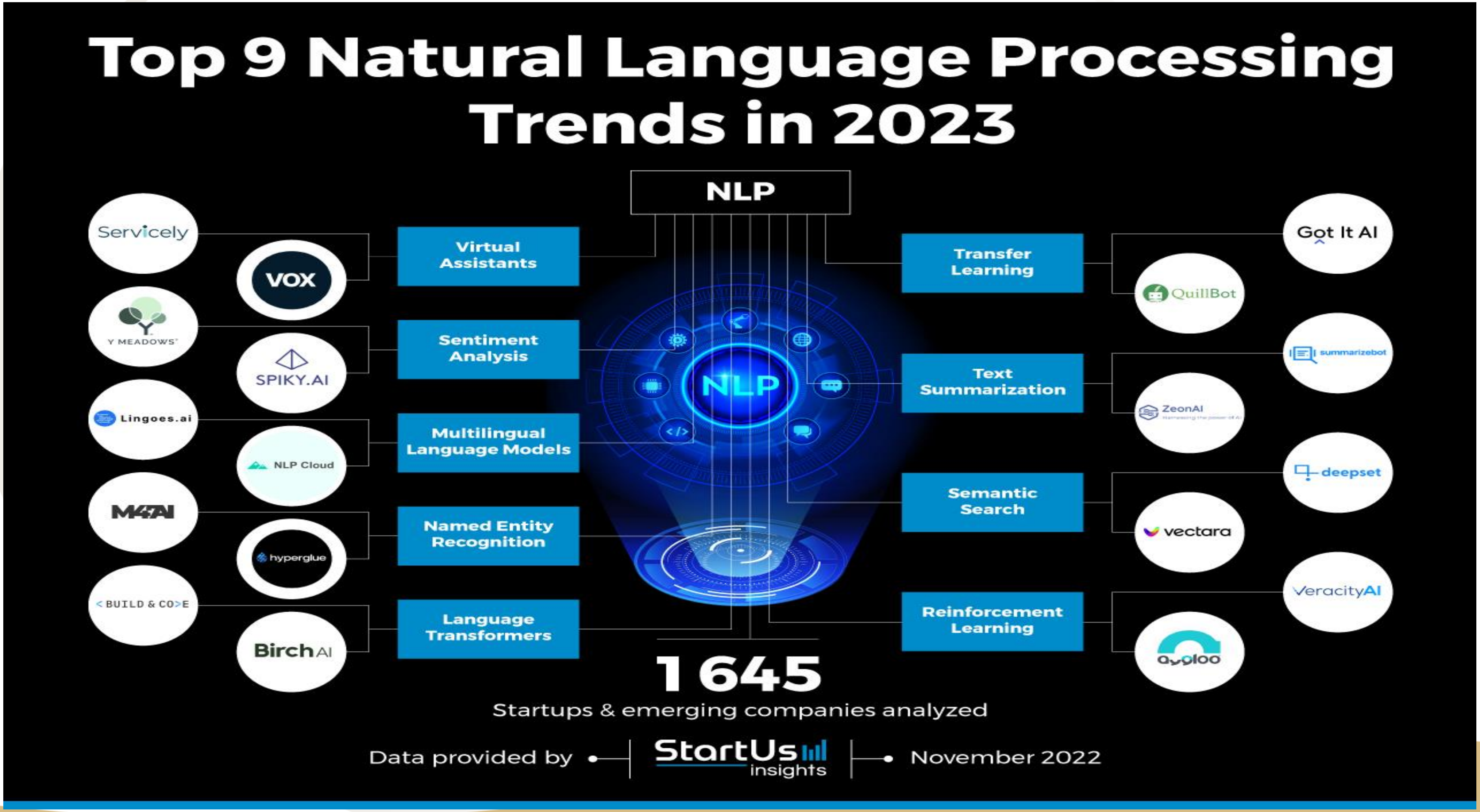
MACHINE LEARNING AND DEEP LEARNING

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.
- Deep learning is a subfield of machine learning focusing on learning data representations as successive layers of increasingly meaningful representations.
- Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.
- Source https://www.ibm.com/topics/machine-learning

APPLICATIONS OF NATURAL LANGUAGE PROCESSING



NLP APPLICATIONS



Top 9 Natural Language Processing Trends in 2023

NLP

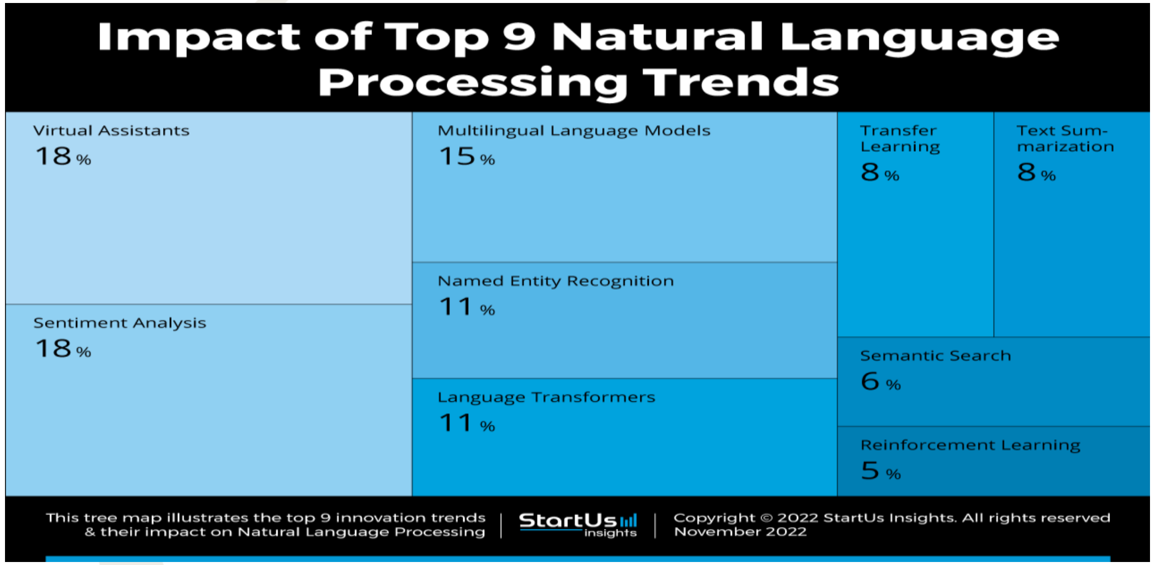
- Virtual Assistants
- Sentiment Analysis
- Multilingual Language Models
- Named Entity Recognition
- Language Transformers
- Transfer Learning
- Text Summarization
- Semantic Search
- Reinforcement Learning

1 645

Startups & emerging companies analyzed

Data provided by StartUs insights November 2022

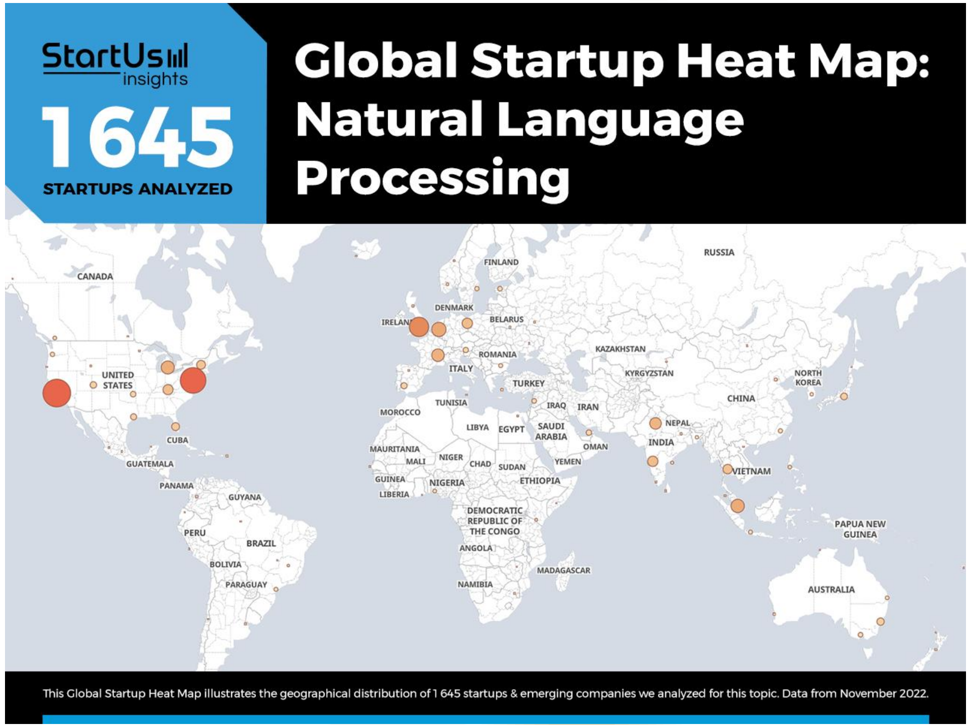
Global Trends in NLP



Impact of Top 9 Natural Language Processing Trends

This tree map illustrates the top 9 innovation trends & their impact on Natural Language Processing | StartUs insights | Copyright © 2022 StartUs Insights. All rights reserved November 2022

https://www.startus-insights.com/innovators-guide/natural-language-processing-trends/



StartUs insights

1645 STARTUPS ANALYZED

Global Startup Heat Map: Natural Language Processing

This Global Startup Heat Map illustrates the geographical distribution of 1 645 startups & emerging companies we analyzed for this topic. Data from November 2022.

NLP APPLICATIONS

- **Sentiment Analysis**
- Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. Today, companies have large volumes of text data like emails, customer support chat transcripts, social media comments, and reviews. Sentiment analysis could be used for more African languges.
- **Text analytics**
- Text analytics converts unstructured text data into meaningful data for analysis using different linguistic, statistical, and machine learning techniques.

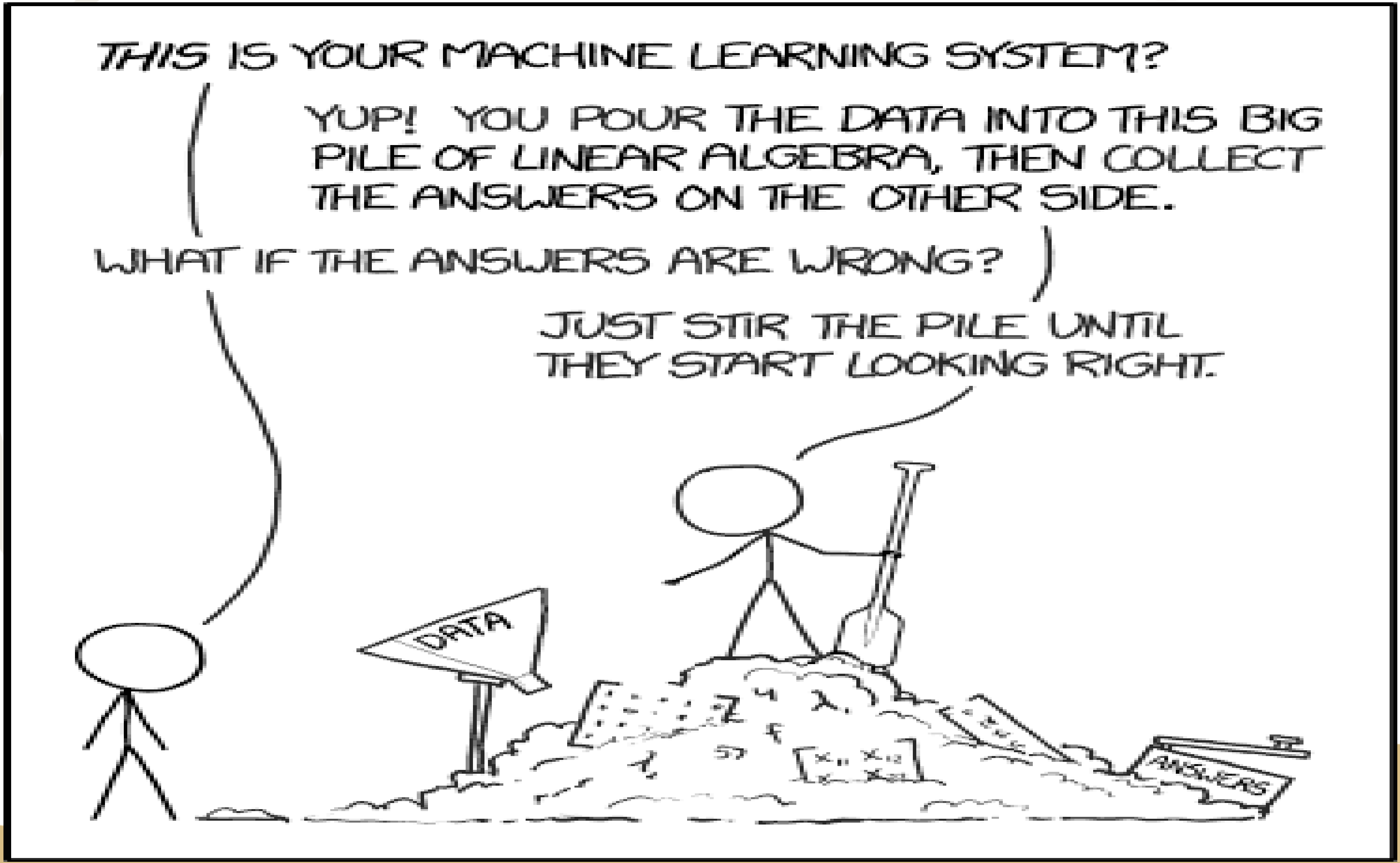
NLP APPLICATIONS

- **Data analysis**
- Natural language capabilities are being integrated into data analysis workflows as more BI vendors offer a natural language interface to data visualizations.
- **Language translation**
- With NLP, online translators can translate languages more accurately and present grammatically-correct results.

NLP APPLICATIONS

- **Predictive text**
- Things like autocorrect, autocomplete, and predictive text are so commonplace on our smartphones that we take them for granted. Autocomplete and predictive text are similar to search engines in that they predict things to say based on what you type, finishing the word or suggesting a relevant one.

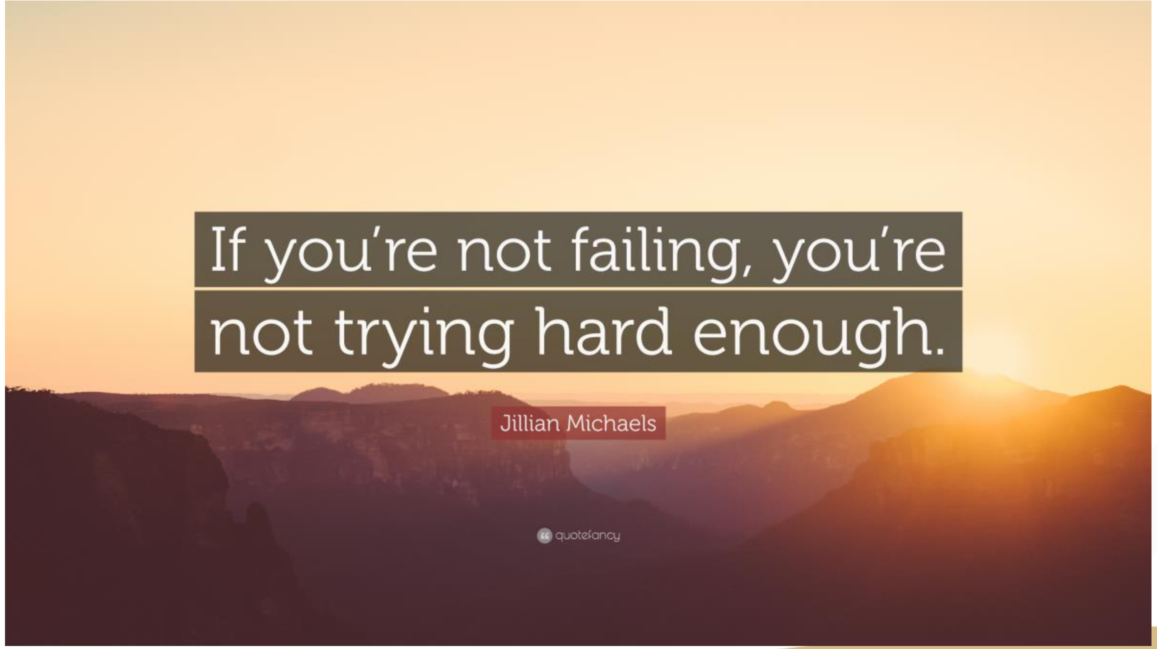
How does it work?


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

NLP HOW DOES IT WORK?

- Identify the structure and meaning of words, sentences, texts and conversations, Deep understanding of broad language
- Tokenization. This is when text is broken down into smaller units to work with.
- 'As', ' she', ' said', ' this', ',', ' she', ' looked', ' down', ' at', ' her', ' hands', ',', ' and', ' was', ' surprised', ' to', ' find', ' that', ' she', ' had', ' put', ' on', ' one', ' of', ' the', ' rabbit', "'s", ' little', '
- Stop word removal. This is when common words are removed from text so unique words that offer the most information about the text remain.
- Lemmatization and stemming. This is when words are reduced to their root forms to process.

A REVIEW OF NLP IN AFRICA…



If you're not failing, you're not trying hard enough.

Jillian Michaels

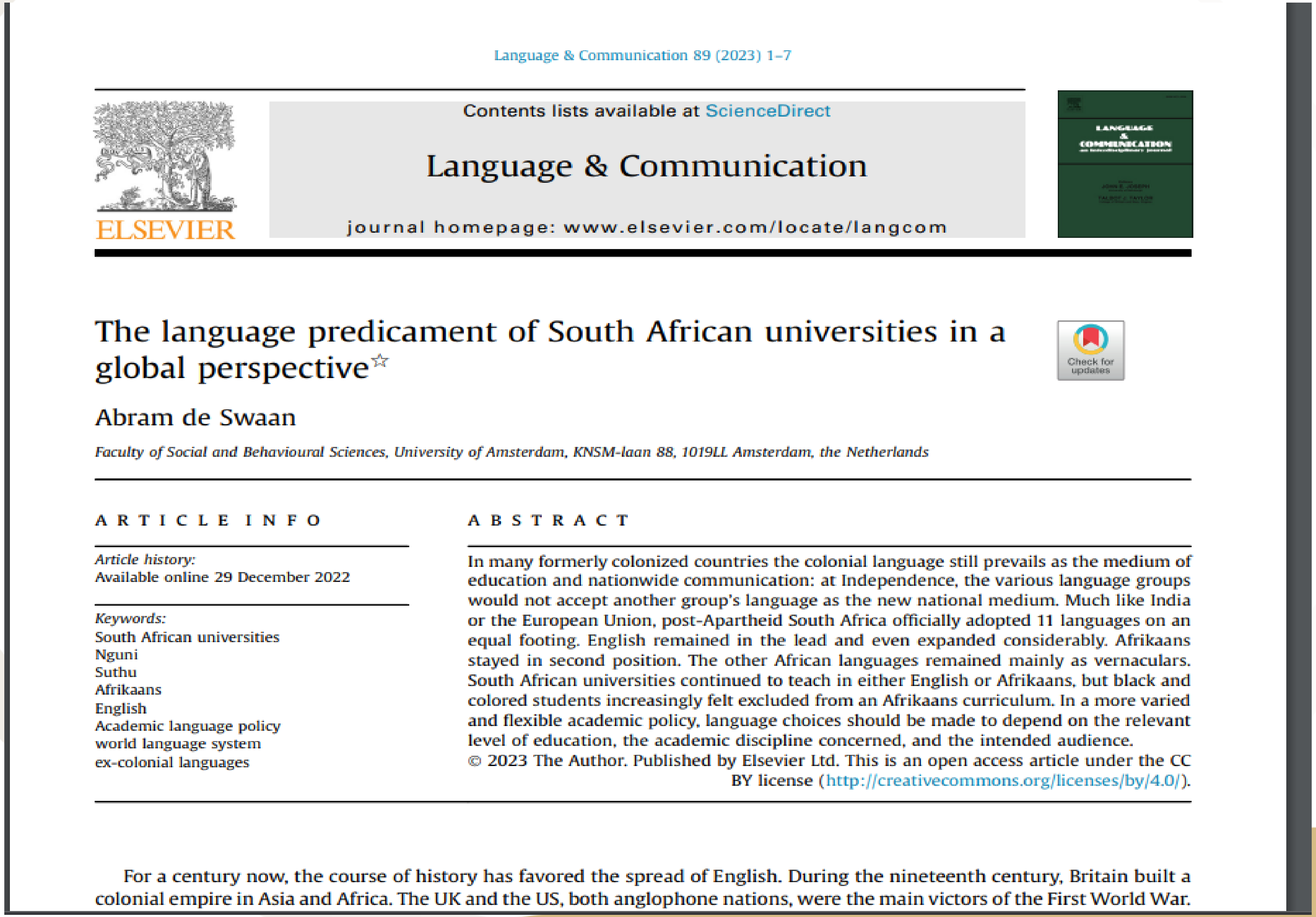
AFRICAN CONTEXT IN NLP

- “Over 1 billion people live in Africa, and its residents speak more than 2,000 languages. But those languages are among the least represented in NLP research, and work on African languages is often side-lined at major venues. In 2022, the wave of large language models built through collaborative networks and large investments in compute has come to the shores of African languages. This year has seen the release of large multilingual models such as BLOOM and NLLB-200 for machine translation. While those models have been publicly open-sourced, their impact on the community of African NLP researchers is yet to be assessed and deserves to be a matter of wider discussion”.
- https://sites.google.com/view/africanlp2023/home

African context: How do we compare?

- There are over 7000 languages spoken around the globe, most NLP processes only use seven languages: English, Chinese, Urdu, Farsi, Arabic, French, and Spanish.

Review on NLP language in Africa.



Language & Communication 89 (2023) 1–7

Contents lists available at ScienceDirect

ELSEVIER

Language & Communication

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

The language predicament of South African universities in a global perspective☆

Abram de Swaan

Faculty of Social and Behavioural Sciences, University of Amsterdam, KNSM-laan 88, 1019LL Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Available online 29 December 2022

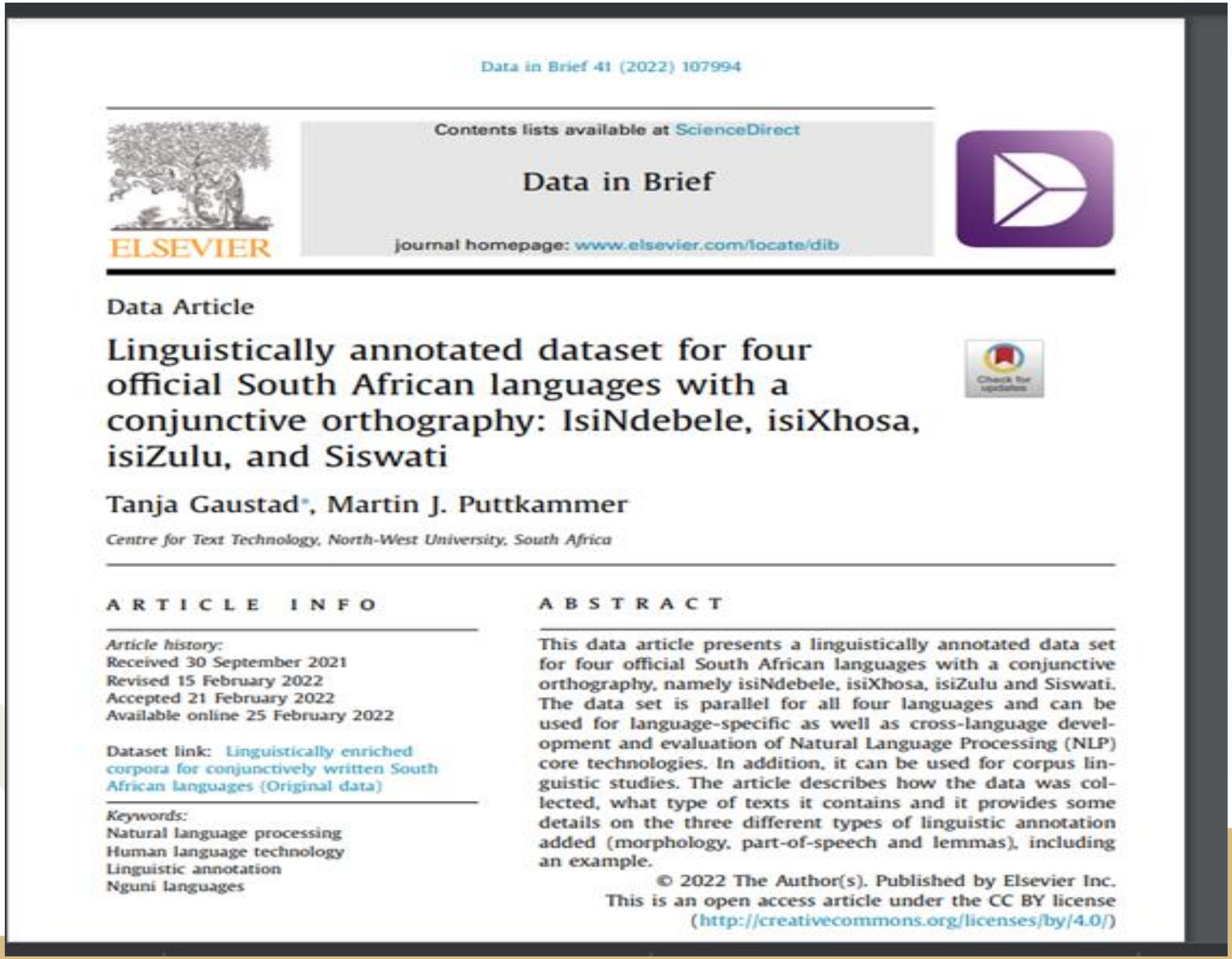
Keywords:
South African universities
Nguni
Suthu
Afrikaans
English
Academic language policy
world language system
ex-colonial languages

ABSTRACT

In many formerly colonized countries the colonial language still prevails as the medium of education and nationwide communication: at Independence, the various language groups would not accept another group's language as the new national medium. Much like India or the European Union, post-Apartheid South Africa officially adopted 11 languages on an equal footing. English remained in the lead and even expanded considerably. Afrikaans stayed in second position. The other African languages remained mainly as vernaculars. South African universities continued to teach in either English or Afrikaans, but black and colored students increasingly felt excluded from an Afrikaans curriculum. In a more varied and flexible academic policy, language choices should be made to depend on the relevant level of education, the academic discipline concerned, and the intended audience.

For a century now, the course of history has favored the spread of English. During the nineteenth century, Britain built a colonial empire in Asia and Africa. The UK and the US, both anglophone nations, were the main victors of the First World War.

Related work on NLP IN AFRICA



Data in Brief 41 (2022) 107994

Contents lists available at ScienceDirect

Data in Brief

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

ELSEVIER

Data Article

Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati

Tanja Gaustad*, Martin J. Puttkammer

Centre for Text Technology, North-West University, South Africa

ARTICLE INFO

Article history:
Received 30 September 2021
Revised 15 February 2022
Accepted 21 February 2022
Available online 25 February 2022

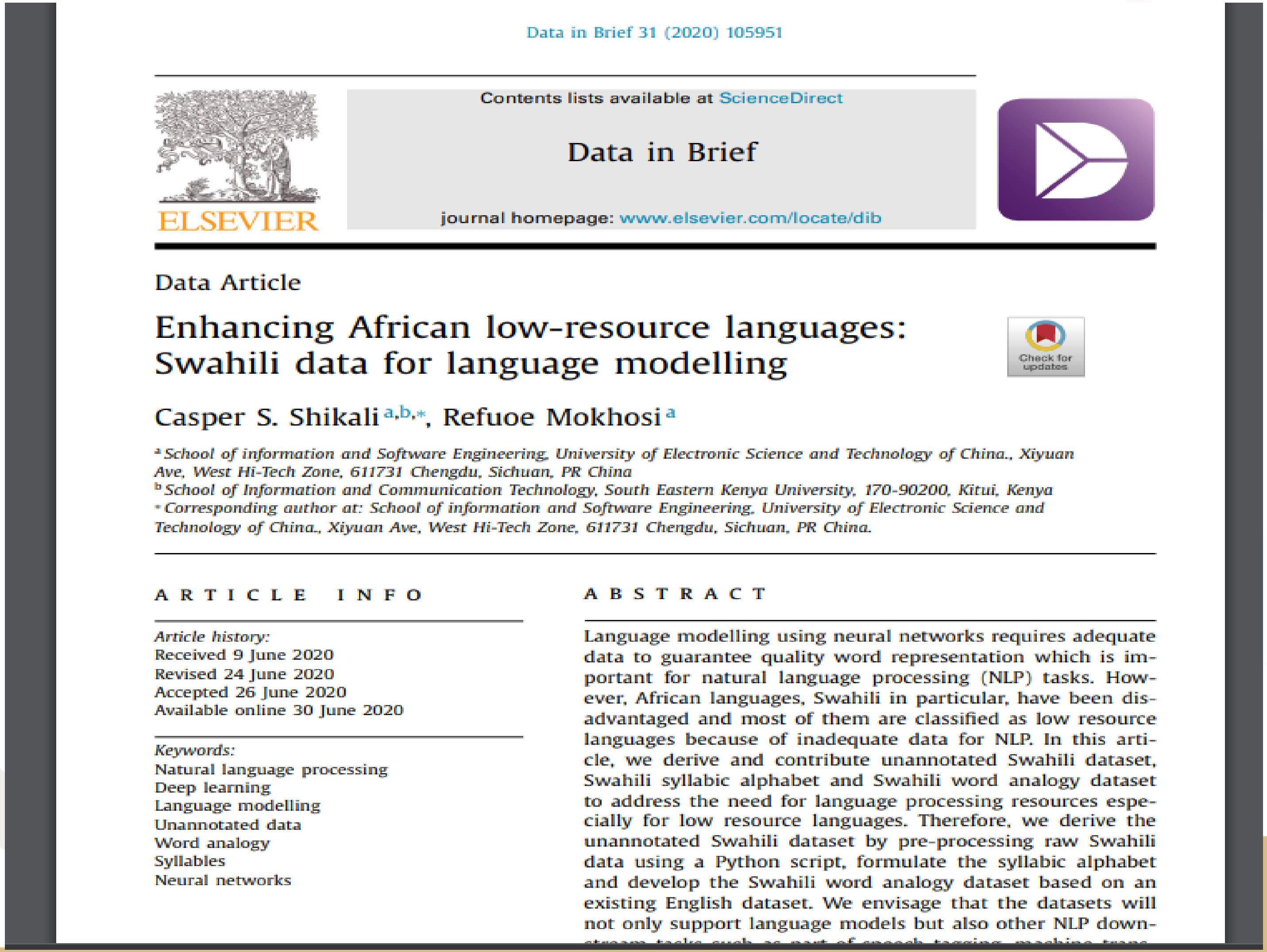
Dataset link: Linguistically enriched corpora for conjunctively written South African languages (Original data)

Keywords:
Natural language processing
Human language technology
Linguistic annotation
Nguni languages

ABSTRACT

This data article presents a linguistically annotated data set for four official South African languages with a conjunctive orthography, namely isiNdebele, isiXhosa, isiZulu and Siswati. The data set is parallel for all four languages and can be used for language-specific as well as cross-language development and evaluation of Natural Language Processing (NLP) core technologies. In addition, it can be used for corpus linguistic studies. The article describes how the data was collected, what type of texts it contains and it provides some details on the three different types of linguistic annotation added (morphology, part-of-speech and lemmas), including an example.

Related work on NLP IN AFRICA



Data in Brief 31 (2020) 105951

Contents lists available at ScienceDirect

Data in Brief

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

ELSEVIER

Data Article

Enhancing African low-resource languages: Swahili data for language modelling

Casper S. Shikali[a,b,*], Refuoe Mokhosi[a]

[a] School of information and Software Engineering, University of Electronic Science and Technology of China., Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, Sichuan, PR China

[b] School of Information and Communication Technology, South Eastern Kenya University, 170-90200, Kitui, Kenya

[*] Corresponding author at: School of information and Software Engineering, University of Electronic Science and Technology of China., Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, Sichuan, PR China.

ARTICLE INFO

Article history:
Received 9 June 2020
Revised 24 June 2020
Accepted 26 June 2020
Available online 30 June 2020

Keywords:
Natural language processing
Deep learning
Language modelling
Unannotated data
Word analogy
Syllables
Neural networks

ABSTRACT

Language modelling using neural networks requires adequate data to guarantee quality word representation which is important for natural language processing (NLP) tasks. However, African languages, Swahili in particular, have been disadvantaged and most of them are classified as low resource languages because of inadequate data for NLP. In this article, we derive and contribute unannotated Swahili dataset, Swahili syllabic alphabet and Swahili word analogy dataset to address the need for language processing resources especially for low resource languages. Therefore, we derive the unannotated Swahili dataset by pre-processing raw Swahili data using a Python script, formulate the syllabic alphabet and develop the Swahili word analogy dataset based on an existing English dataset. We envisage that the datasets will not only support language models but also other NLP downstream tasks such as part of speech tagging, machine trans-

Problems associated with NLP in an African context

Even in the forums which aim to widen NLP participation, Africa is barely represented - despite the fact that Africa has over 2000 languages. The 4th Industrial revolution in Africa cannot take place in English. It is imperative that NLP models be developed for the African continent

- As per Martinus (2019), some problems facing NLP in African languages are as follows:

•**Focus**: According to Alexander (2009), African society does not see hope for indigenous languages to be accepted as a more primary mode for communication. As a result, there are few efforts to fund and focus on support of these languages, despite their potential impact

•**Low Resourced:** The lack of resources for African languages hinders the ability for researchers to do NLP

•**Low Discoverability:** The resources for African languages that do exist are hard to find. Often one needs to be associated with a specific academic institution in a specific country to gain access to the language data available for that country. This reduces the ability of countries and institutions to combine their knowledge and datasets to achieve better performance and innovations. Often the existing research itself is hard to discover since they are often published in smaller African conferences or journals, which are not electronically available nor indexed by research tools such as Google Scholar.

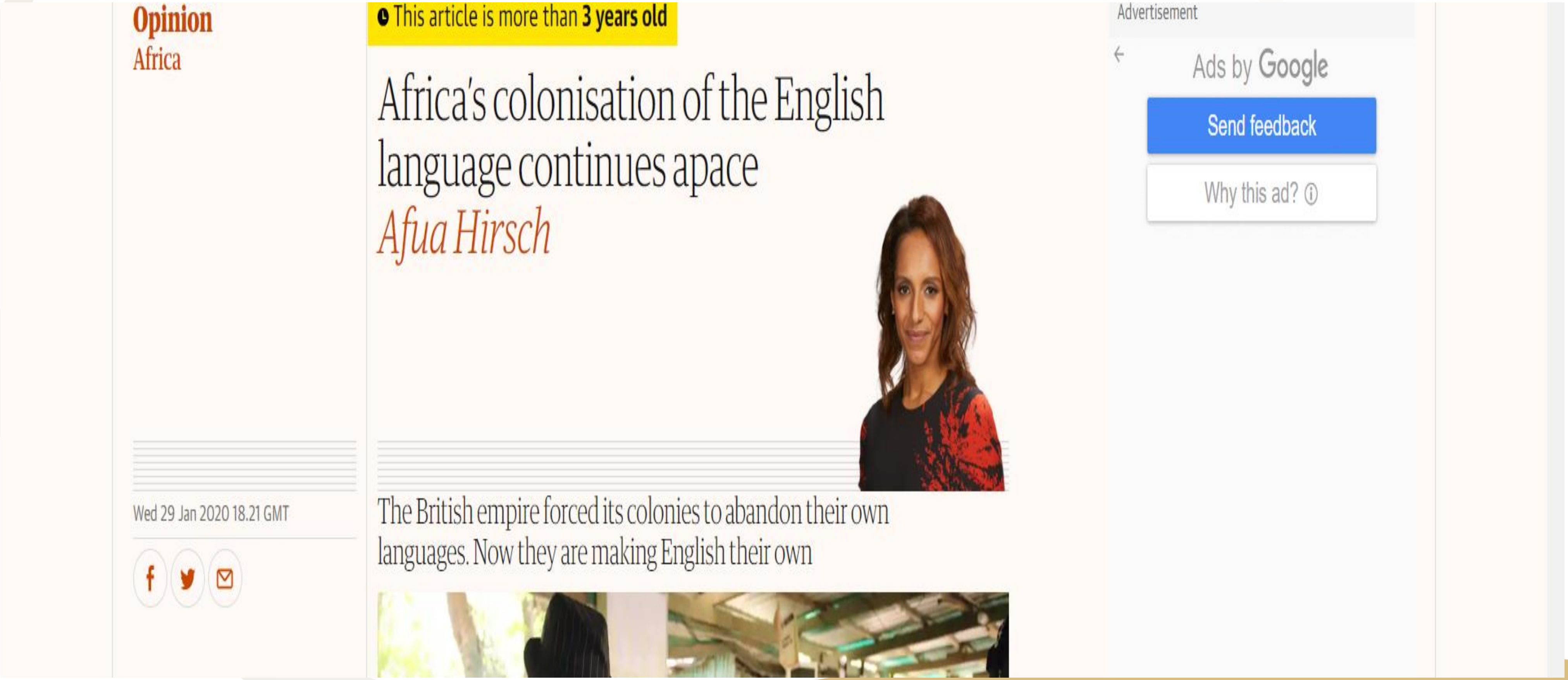
•**Lack of publicly**-available benchmarks: Due to the low discoverability and the lack of research in the field, there are no publicly available benchmarks or leaderboards to new compare NLP techniques to

•**Reproducibility:** The data and code of existing research are rarely shared, which means researchers cannot reproduce the results properly.

What is Language colonisation

- *“The domination of a people’s language by the languages of the colonising nations was crucial to the domination of the mental universe of the colonised.” Ngũgĩ wa Thiong’o*

LANGUAGE COLONISATION



Opinion
Africa

This article is more than 3 years old

Africa's colonisation of the English language continues apace

Afua Hirsch

Wed 29 Jan 2020 18.21 GMT

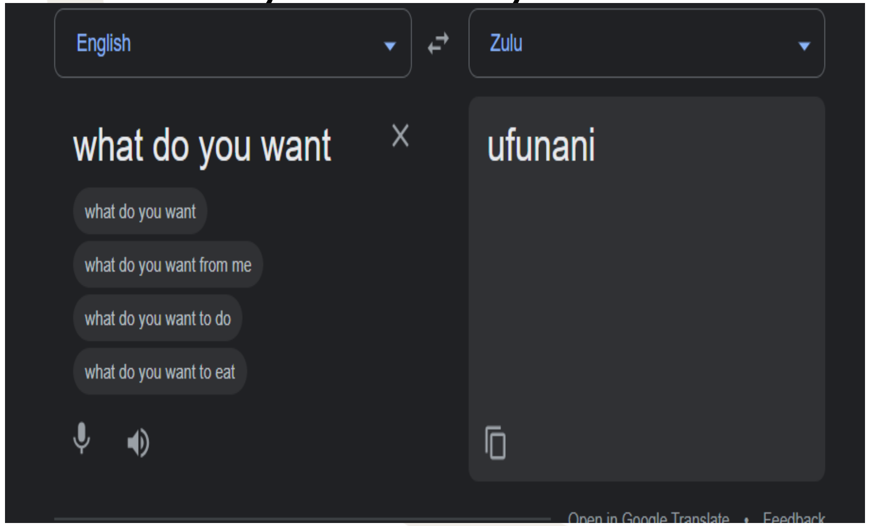
The British empire forced its colonies to abandon their own languages. Now they are making English their own

Language colonisation: how did it happen

- Most words in Africa were borrowed more from Eurocentric languages.
- Bicycle → Mbasikolo
- English names were given on the premise of western world failing to pronounce African names.
- As pronunciation of names and words became difficult to pronounce for the coloniser, substitutes of so-called “Christian” names were given to Africans

Language colonisation in NLP context

- Lack of translators for most African languages to Western languages.
- Lack of sentiment analysers on African languages. e.g. on twitter, tiktok, facebook, Chatbox, teams, WhatsApp.



English ⇄ Zulu

what do you want

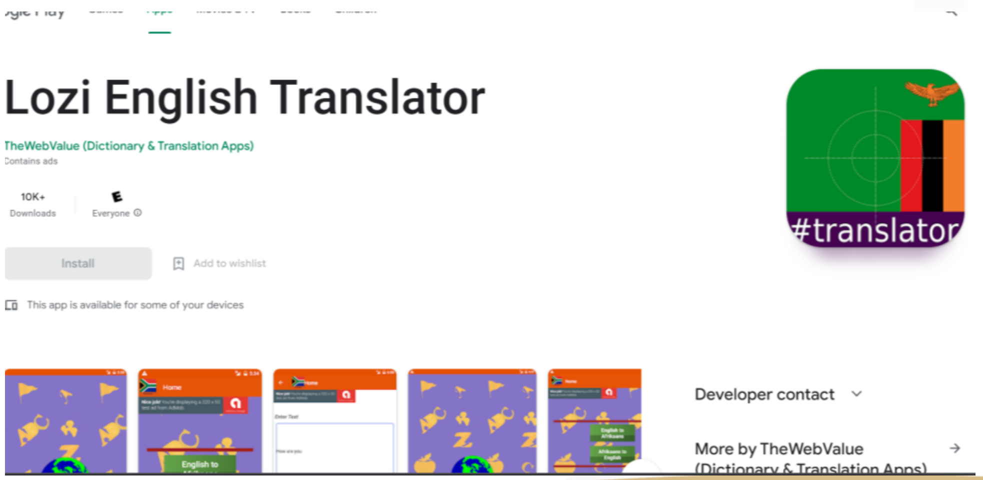
ufunani

what do you want

what do you want from me

what do you want to do

what do you want to eat



Lozi English Translator

TheWebValue (Dictionary & Translation Apps)

Contains ads

10K+ Downloads

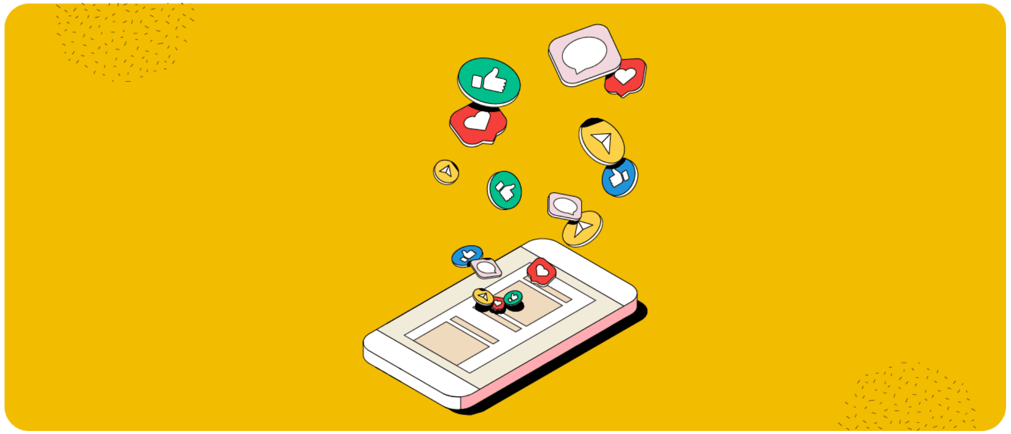
E Everyone

This app is available for some of your devices

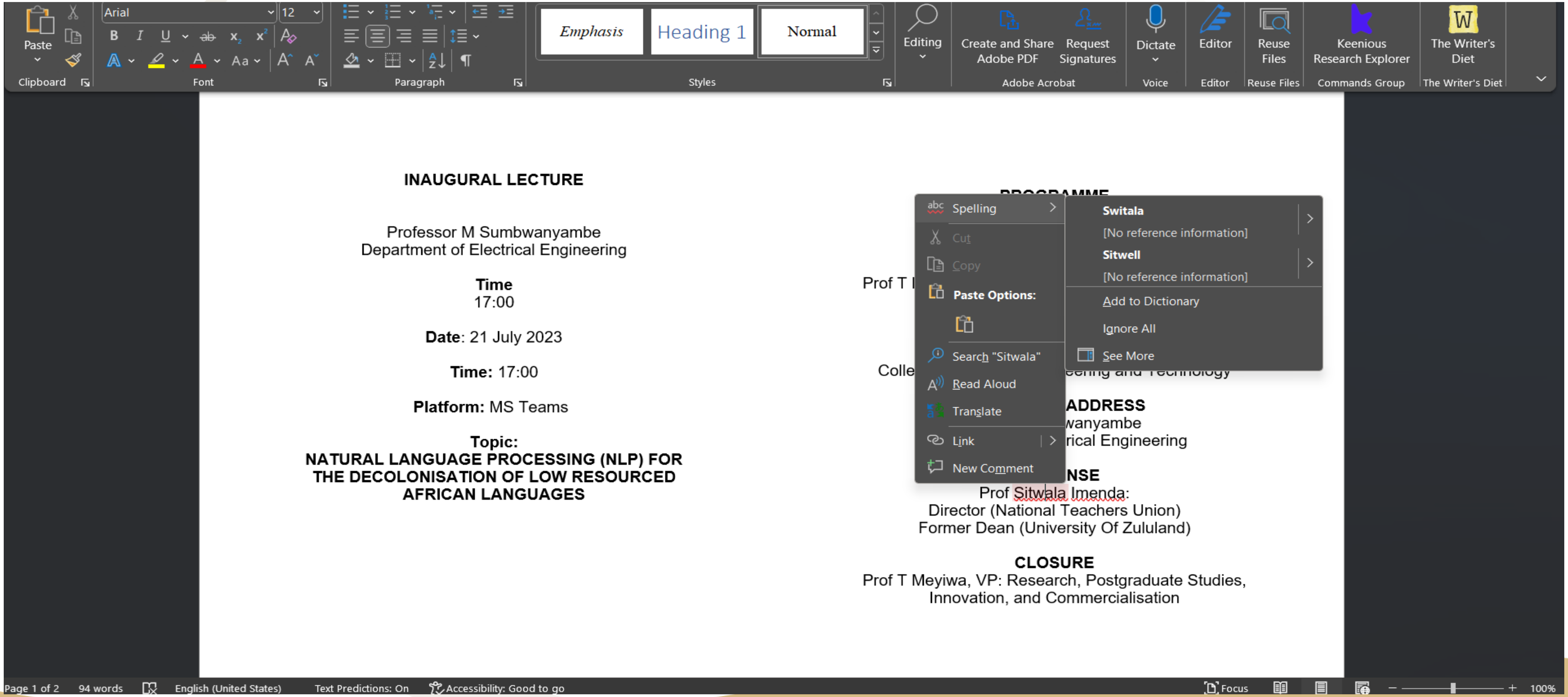
More by TheWebValue (Dictionary & Translation Apps)

The culprits

- The culprits include most of the social media platforms where translation is done for most common languages but not for African languages.



How do we decolonise African languages through NLP?



INAUGURAL LECTURE

Professor M Sumbwanyambe
Department of Electrical Engineering

Time
17:00

Date: 21 July 2023

Time: 17:00

Platform: MS Teams

Topic:
NATURAL LANGUAGE PROCESSING (NLP) FOR THE DECOLONISATION OF LOW RESOURCED AFRICAN LANGUAGES

PROGRAMME

Prof T I

Colle

eering and Technology

ADDRESS
wanyambe
rical Engineering

NSE
Prof Sitwala Imenda:
Director (National Teachers Union)
Former Dean (University Of Zululand)

CLOSURE
Prof T Meyiwa, VP: Research, Postgraduate Studies, Innovation, and Commercialisation

- THE TIME TO DECOLONISE OUR LANGUGES IS NOW
- AND HOW DO WE DO IT?

- WE NEED TO DECOLONISE OUR TECHNOLOGY

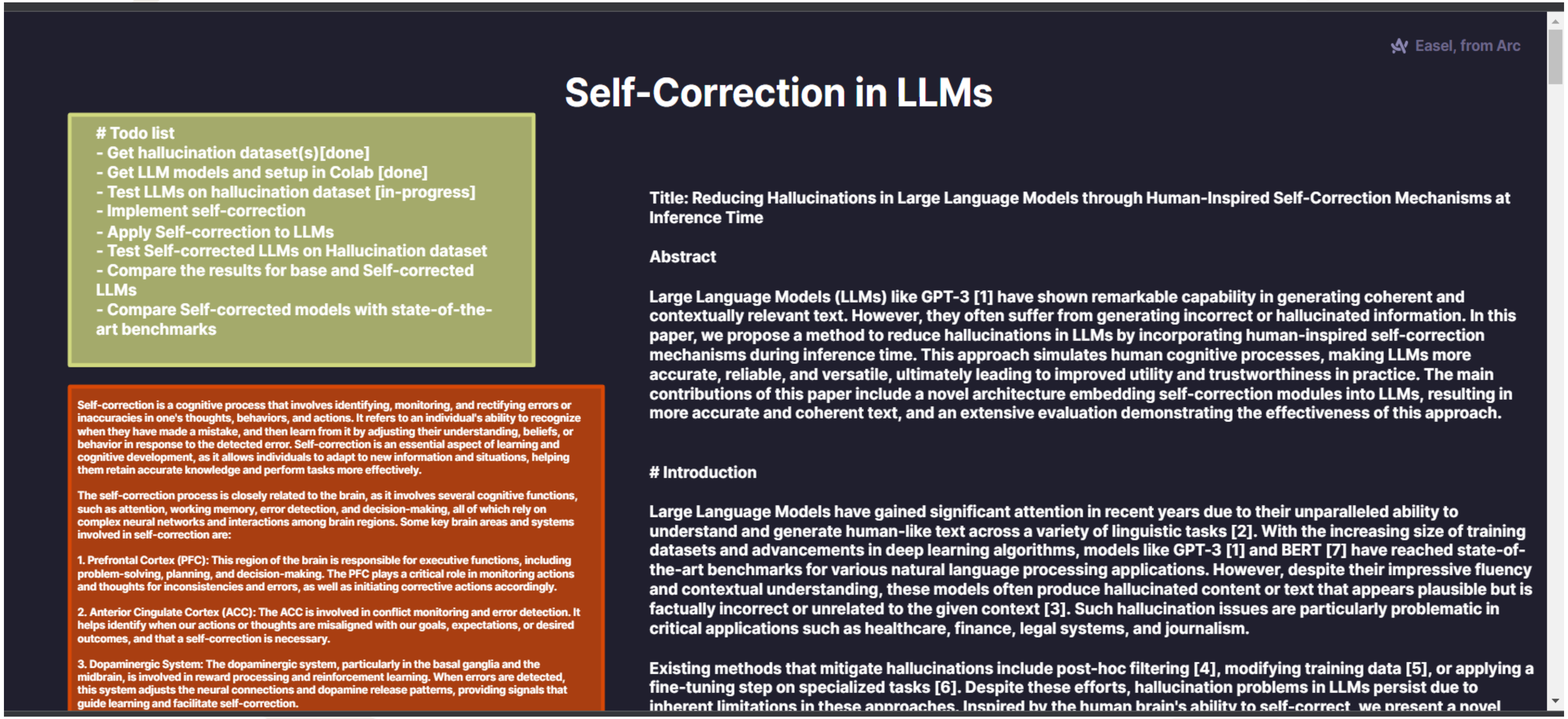
Large Language Models (LLM)

- Large Language Models have gained significant attention in recent years due to their unparalleled ability to understand and generate human-like text across a variety of linguistic tasks. With the increasing size of training datasets and advancements in deep learning algorithms, models like GPT-3 and Bidirectional Encoder Representations from Transformers (BERT) have reached state-of-the-art benchmarks for various natural language processing applications.
- For example, the GPT-3 model that is backing the ChatGPT service was trained on massive amounts of text data from the internet. This includes books, articles, websites, and various other sources. During the training process, the model learns the statistical relationships between words, phrases, and sentences, allowing it to generate coherent and contextually relevant responses when given a prompt or query

LLM

- These have not been trained on African corpus or names.
- Lack of huge African datasets prevents the training of the LLM on African languages.

CURRENT RESEARCH



Self-Correction in LLMs

Todo list
- Get hallucination dataset(s)[done]
- Get LLM models and setup in Colab [done]
- Test LLMs on hallucination dataset [in-progress]
- Implement self-correction
- Apply Self-correction to LLMs
- Test Self-corrected LLMs on Hallucination dataset
- Compare the results for base and Self-corrected LLMs
- Compare Self-corrected models with state-of-the-art benchmarks

Self-correction is a cognitive process that involves identifying, monitoring, and rectifying errors or inaccuracies in one's thoughts, behaviors, and actions. It refers to an individual's ability to recognize when they have made a mistake, and then learn from it by adjusting their understanding, beliefs, or behavior in response to the detected error. Self-correction is an essential aspect of learning and cognitive development, as it allows individuals to adapt to new information and situations, helping them retain accurate knowledge and perform tasks more effectively.

The self-correction process is closely related to the brain, as it involves several cognitive functions, such as attention, working memory, error detection, and decision-making, all of which rely on complex neural networks and interactions among brain regions. Some key brain areas and systems involved in self-correction are:

1. Prefrontal Cortex (PFC): This region of the brain is responsible for executive functions, including problem-solving, planning, and decision-making. The PFC plays a critical role in monitoring actions and thoughts for inconsistencies and errors, as well as initiating corrective actions accordingly.

2. Anterior Cingulate Cortex (ACC): The ACC is involved in conflict monitoring and error detection. It helps identify when our actions or thoughts are misaligned with our goals, expectations, or desired outcomes, and that a self-correction is necessary.

3. Dopaminergic System: The dopaminergic system, particularly in the basal ganglia and the midbrain, is involved in reward processing and reinforcement learning. When errors are detected, this system adjusts the neural connections and dopamine release patterns, providing signals that guide learning and facilitate self-correction.

Title: Reducing Hallucinations in Large Language Models through Human-Inspired Self-Correction Mechanisms at Inference Time

Abstract

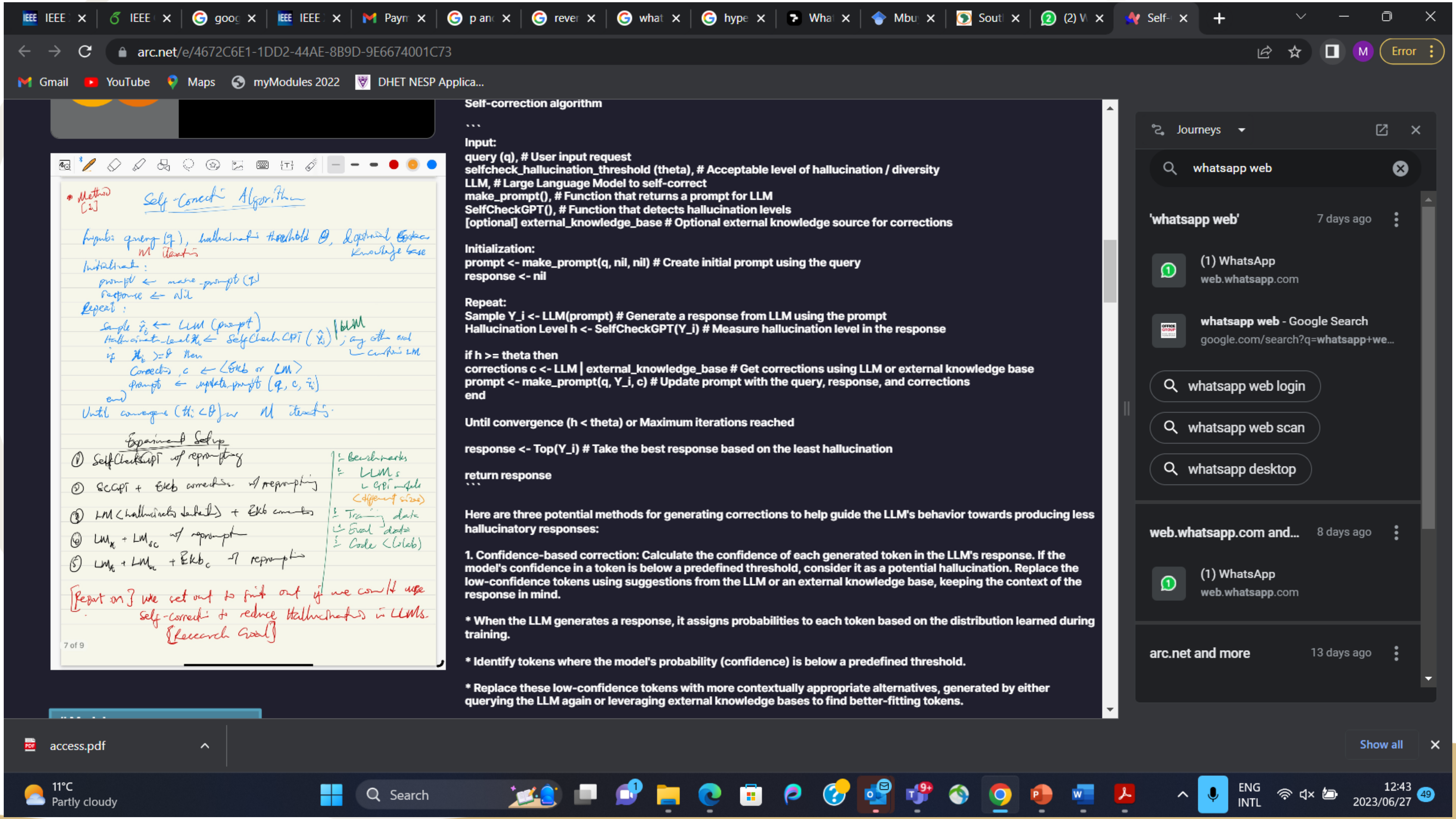
Large Language Models (LLMs) like GPT-3 [1] have shown remarkable capability in generating coherent and contextually relevant text. However, they often suffer from generating incorrect or hallucinated information. In this paper, we propose a method to reduce hallucinations in LLMs by incorporating human-inspired self-correction mechanisms during inference time. This approach simulates human cognitive processes, making LLMs more accurate, reliable, and versatile, ultimately leading to improved utility and trustworthiness in practice. The main contributions of this paper include a novel architecture embedding self-correction modules into LLMs, resulting in more accurate and coherent text, and an extensive evaluation demonstrating the effectiveness of this approach.

Introduction

Large Language Models have gained significant attention in recent years due to their unparalleled ability to understand and generate human-like text across a variety of linguistic tasks [2]. With the increasing size of training datasets and advancements in deep learning algorithms, models like GPT-3 [1] and BERT [7] have reached state-of-the-art benchmarks for various natural language processing applications. However, despite their impressive fluency and contextual understanding, these models often produce hallucinated content or text that appears plausible but is factually incorrect or unrelated to the given context [3]. Such hallucination issues are particularly problematic in critical applications such as healthcare, finance, legal systems, and journalism.

Existing methods that mitigate hallucinations include post-hoc filtering [4], modifying training data [5], or applying a fine-tuning step on specialized tasks [6]. Despite these efforts, hallucination problems in LLMs persist due to inherent limitations in these approaches. Inspired by the human brain's ability to self-correct, we present a novel

Current research in NLP and LLM



* Method [1]

Self-Correction Algorithm

```
Input: query (q), hallucination threshold θ, [optional External Knowledge base]
       M iterations
Initialization:
    prompt ← make_prompt(q)
    response ← nil
Repeat:
    Sample $\hat{Y}_i$ ← LLM (prompt)
    Hallucination_level $H_i$ ← SelfCheckGPT($\hat{Y}_i$)   | LLM ; any other eval
                                                           ↳ confirm LM
    if $H_i$ >= θ then
        Corrections, c ← <Ekb or LM>
        prompt ← update_prompt(q, c, $\hat{Y}_i$)
    end
Until convergence ($H_i$ < θ) or M iterations.
```

Experiment Setup

1. SelfCheckGPT w/ reprompting
2. SCGPT + Ekb corrections. w/ reprompting
3. LM <hallucination detector> + Ekb corrections
4. $LM_R + LM_{SC}$ w/ reprompt
5. $LM_R + LM_{SC} + Ekb_c$ w/ reprompting

- Benchmarks
- LLMs
  ↳ GPT models <different sizes>
- Training data
  ↳ Eval data
- Code <Colab>

[Report on] we set out to find out if we could use self-correction to reduce Hallucinations in LLMs. [Research Goal]

7 of 9

Self-correction algorithm

```
Input:
query (q), # User input request
selfcheck_hallucination_threshold (theta), # Acceptable level of hallucination / diversity
LLM, # Large Language Model to self-correct
make_prompt(), # Function that returns a prompt for LLM
SelfCheckGPT(), # Function that detects hallucination levels
[optional] external_knowledge_base # Optional external knowledge source for corrections

Initialization:
prompt <- make_prompt(q, nil, nil) # Create initial prompt using the query
response <- nil

Repeat:
Sample Y_i <- LLM(prompt) # Generate a response from LLM using the prompt
Hallucination Level h <- SelfCheckGPT(Y_i) # Measure hallucination level in the response

if h >= theta then
corrections c <- LLM | external_knowledge_base # Get corrections using LLM or external knowledge base
prompt <- make_prompt(q, Y_i, c) # Update prompt with the query, response, and corrections
end

Until convergence (h < theta) or Maximum iterations reached

response <- Top(Y_i) # Take the best response based on the least hallucination

return response
```

Here are three potential methods for generating corrections to help guide the LLM's behavior towards producing less hallucinatory responses:

1. Confidence-based correction: Calculate the confidence of each generated token in the LLM's response. If the model's confidence in a token is below a predefined threshold, consider it as a potential hallucination. Replace the low-confidence tokens using suggestions from the LLM or an external knowledge base, keeping the context of the response in mind.

* When the LLM generates a response, it assigns probabilities to each token based on the distribution learned during training.

* Identify tokens where the model's probability (confidence) is below a predefined threshold.

* Replace these low-confidence tokens with more contextually appropriate alternatives, generated by either querying the LLM again or leveraging external knowledge bases to find better-fitting tokens.

NECESSARY THINGS TO ACVHIVE THE DECOLONISATION

- Lack of Relevant Government Policies
- Ethics
- User Attitudes
- Insufficient Infrastructure and Network Connectivity
- Lack of Structured Data Ecosystem
- Lack of Skills Acquisition
- Lack of structured corpus in low resourced languages.

NLP AND THE FUTURE OF AFRICAN LANGUAGES

- Research has shown that African languages through lagging behind have a potential to be included in the NLP and LLM applications, but a lot is being done to catch up with the west whose NLP are well advanced.
- The new neural network model, which the researchers have dubbed AfriBERTa, uses deep-learning techniques to achieve state-of-the-art results for low-resource languages.

- The neural network language model works specifically with 11 African languages, such as Amharic, Hausa, and Swahili, spoken collectively by more than 400 million people. It achieves competitive output quality despite learning from just one gigabyte of text, while other models require thousands of times more data.

BENEFITS TOWARDS SOCIAL AND ECONOMIC DEVELOPMENT

- Rural people in Africa will be involved in e-commerce.
- Teaching and learning can be done in one’s language by using translators.

- In essence we cannot decolonise before we decolonise technology.

THANK YOU.