Topic and Sentiment Models for Humanoid Emotions

DR. EMEKA OGBUJU

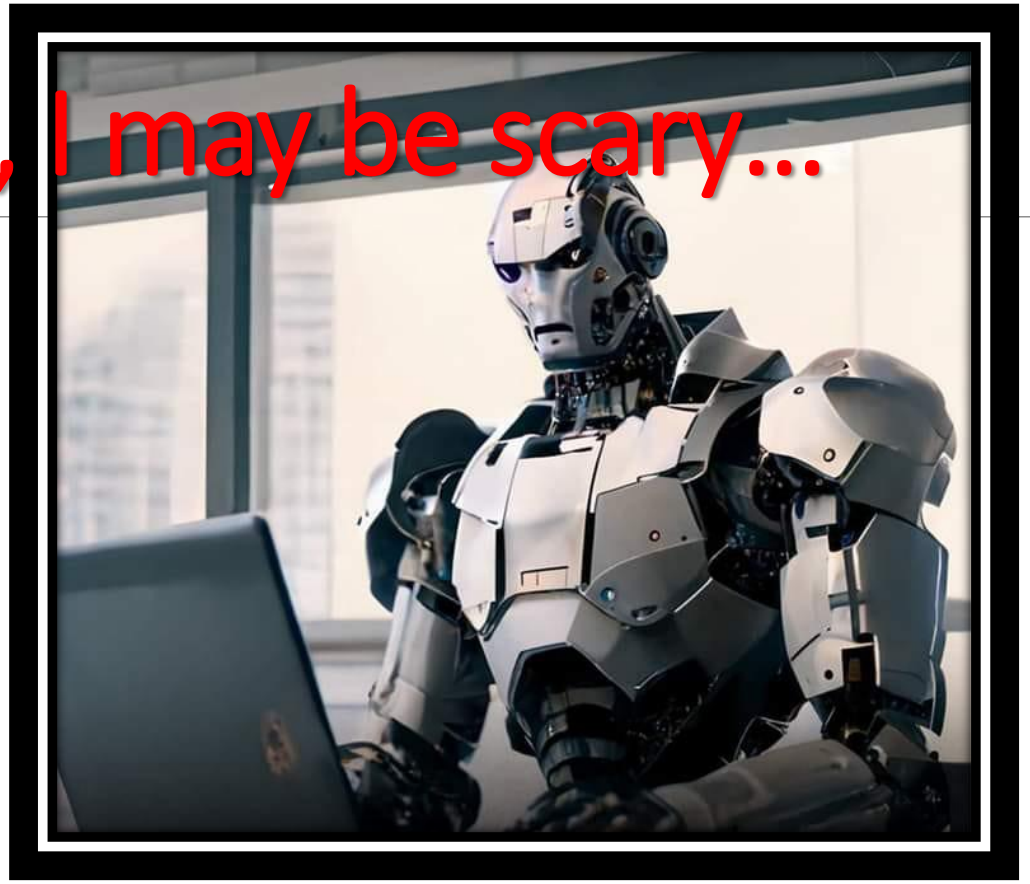
About Me

- B.Sc, M.Sc in Computer Science; PhD in Artificial Intelligence
- Deep Learning Indaba and Black in AI communities in South Africa & Canada
- Senior Lecturer at Federal University Lokoja
- Husband & Father

Presentation Outline

- The importance of understanding humanoid emotions (texts and images) - Overview of Python Programming and R/Studio Libraries for the workshop
- Topic Modeling for Humanoid Emotions Recognition (HER) - LDA and Examples Sentiment Analysis for HER - polarity vs emotional analysis
- Applications - Practical implementation of how topic and sentiment models can be used in HER (Text analytic demos) - Examples of real-world applications of HER - Visualizations/Charts
- Challenges and Future Directions- Current challenges in developing effective topic and sentiment models for HER - Potential future directions for research in this area

Sometimes, I may be scary…



understanding humanoid emotions recognition is critical for improving the interaction between humans and robots and for creating more personalized and effective experiences for users.

Why HER

1. **Improved communication:**

   Humans communicate through both verbal and nonverbal cues. Recognizing emotions expressed through facial expressions, gestures, and other nonverbal cues helps robots and other humanoid devices communicate more effectively with humans.

2. **Enhanced empathy:**

   By recognizing and responding to emotions, humanoid devices can exhibit greater empathy towards humans. This can improve the human-device interaction and make users feel more comfortable and understood.

Why HER

3. **Better user experience:**

   If a humanoid device is able to recognize and respond to emotions, it can provide a more personalized and customized experience for the user. For example, a device that can recognize when a user is feeling stressed may suggest activities to help them relax.

4. **Improved healthcare:**

   Emotion recognition can be particularly useful in healthcare settings, where patients may be unable to communicate their emotions or symptoms verbally. By recognizing emotions and other nonverbal cues, humanoid devices can help healthcare providers better understand and respond to patients' needs.

Why HER

5. **Increased safety:**

   Recognizing emotions can help humanoid devices understand when a human is in distress or experiencing discomfort. This can lead to faster responses and interventions, improving the safety of the user.

Why Topic Modelling

Topic modelling is a technique used to identify patterns in large datasets of text or speech. It discovers the hidden thematic structure in a collection of texts. It can be used to analyze spoken or written content to identify themes or topics that may be associated with certain emotions.

Topic modelling is a powerful tool for understanding human emotions and behaviors, and can be an effective technique for improving humanoid emotion recognition and creating more personalized experiences for users.

Why Topic Modelling

For example, if a humanoid device is designed to recognize when a user is feeling anxious or stressed, it may analyze the user's speech or text messages to identify topics that are commonly associated with anxiety or stress, such as work-related deadlines or personal relationships.

The device may suggest relaxation techniques or mindfulness exercises to help the user manage their stress.

It can also be used to analyze social media content or online reviews to identify patterns in how users talk about different products or services for companies that want to understand how their products are perceived by customers and how they can improve the user experience.

Examples of Topic Models

1. Latent Dirichlet Allocation (LDA): It assumes that each document is a mixture of multiple topics, and each topic is a distribution of words. It aims to discover the latent topics that explain the co-occurrence of words in a document collection.
2. Non-negative Matrix Factorization (NMF): It factorizes the term-document matrix into two non-negative matrices representing the topics and the document-topic distributions. It is widely used in text mining and image analysis.
3. Hierarchical Dirichlet Process (HDP): It is a Bayesian extension of LDA that allows for an infinite number of topics. It is useful in scenarios where the number of topics is unknown or may grow over time.
4. Probabilistic Latent Semantic Analysis (PLSA): It assumes that each document is a mixture of topics, but unlike LDA, it does not model the generative process for the topics themselves.
5. Correlated Topic Model (CTM): CTM is a variant of LDA that allows for correlations between topics, which can capture more complex relationships between topics.

Why Sentiment Analysis

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. – IBM, 2020

Since 80% of data in the world resides in an unstructured format, text mining is an extremely valuable practice within organizations.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Some Emotion Models

Model	Type	Description
Ekman (Ekman, 1992)	Categorical	Six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) characterized by distinctive universal signals, and physiology
Tomkins (Tomkins, 2008)	Categorical	Seven affects organized in low/high intensity couples (interest-excitement, enjoyment-joy, surprise-startle, distress-anguish, anger-rage, fear-terror, shame-humiliation), plus disgust, and *dismell*
Valence/Arousal (VA) (Russell, 1980)	Dimensional	Emotions represented over a two-dimensional circular space: axes describe the valence and arousal
Pleasure/Arousal/Dominance (PAD) (Mehrabian, 1996)	Dimensional	Emotions described with three dimensions: pleasure, arousal, and dominance
3D hypercube (Trnka et al., 2016)	Dimensional	Emotions described with three dimensions: valence, intensity, controllability, and utility
Plutchik wheel of emotions (Plutchik and Kellerman, 2013)	Hybrid	Eight primary emotions differentiated by levels of intensity

For each entry the type of model and a brief description are reported.

Why Sentiment Analysis

Also called as Opinion extraction, Opinion mining, Sentiment mining, Subjectivity analysis

Applications

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

$89 online, **$100** nearby ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

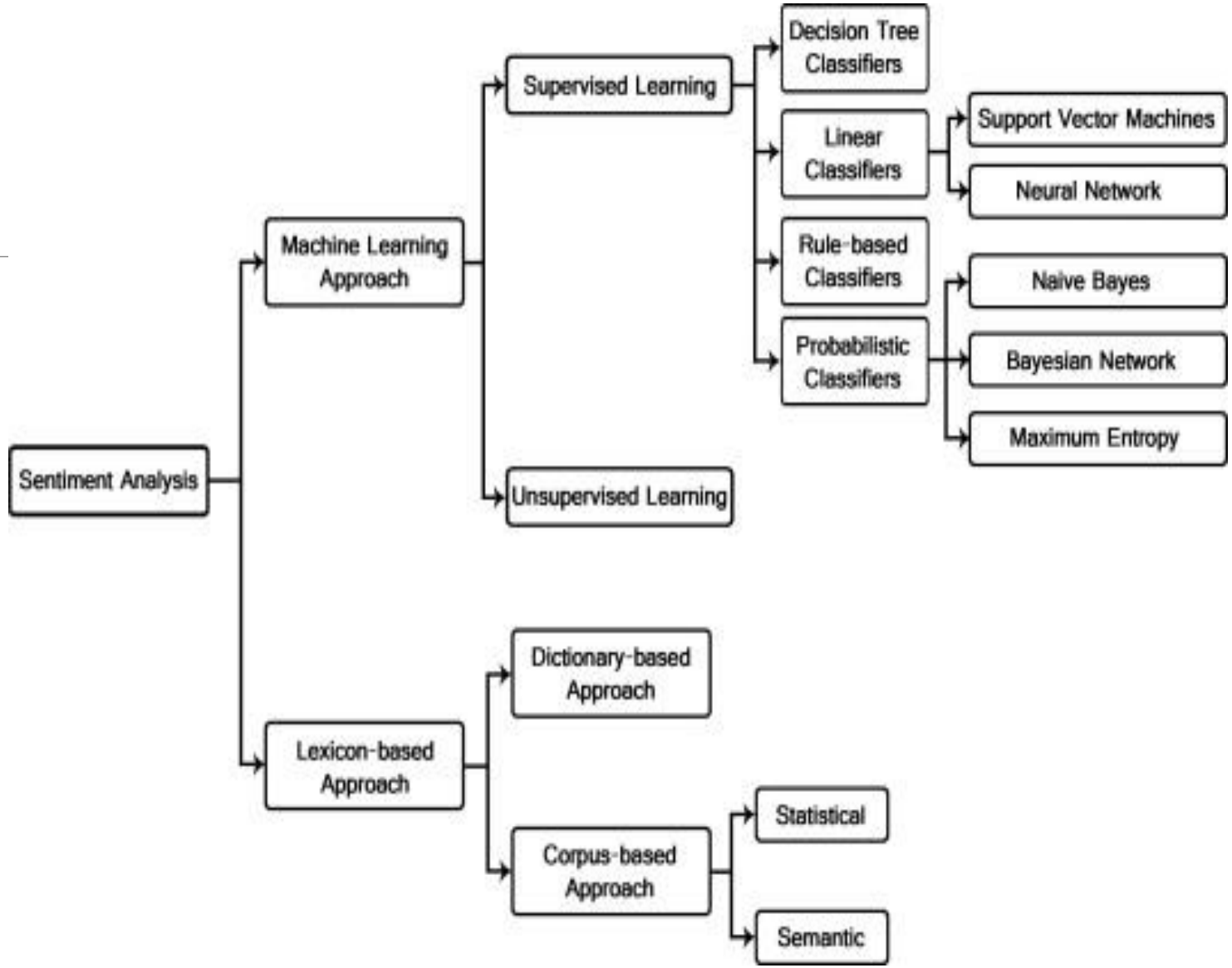
1 star	2	3	4 stars	5 stars

What people are saying

ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

Approaches

1. **ML Approach**
2. **Lexicon Approach**

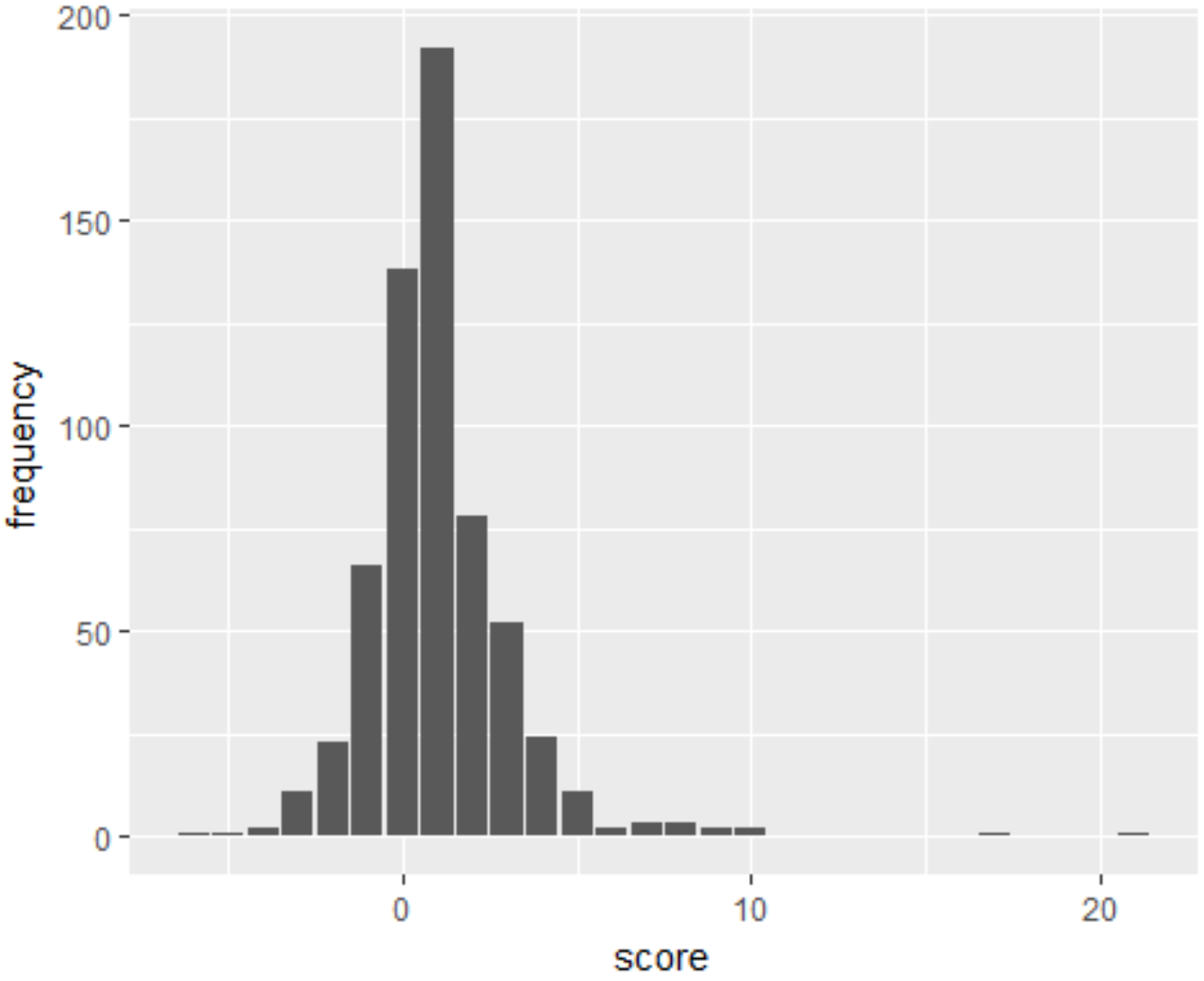


Source: Medhat *et al.*, 2014

Bing Lexicon Approach: Polarity Analysis

2006 positive words and 4783 negative words

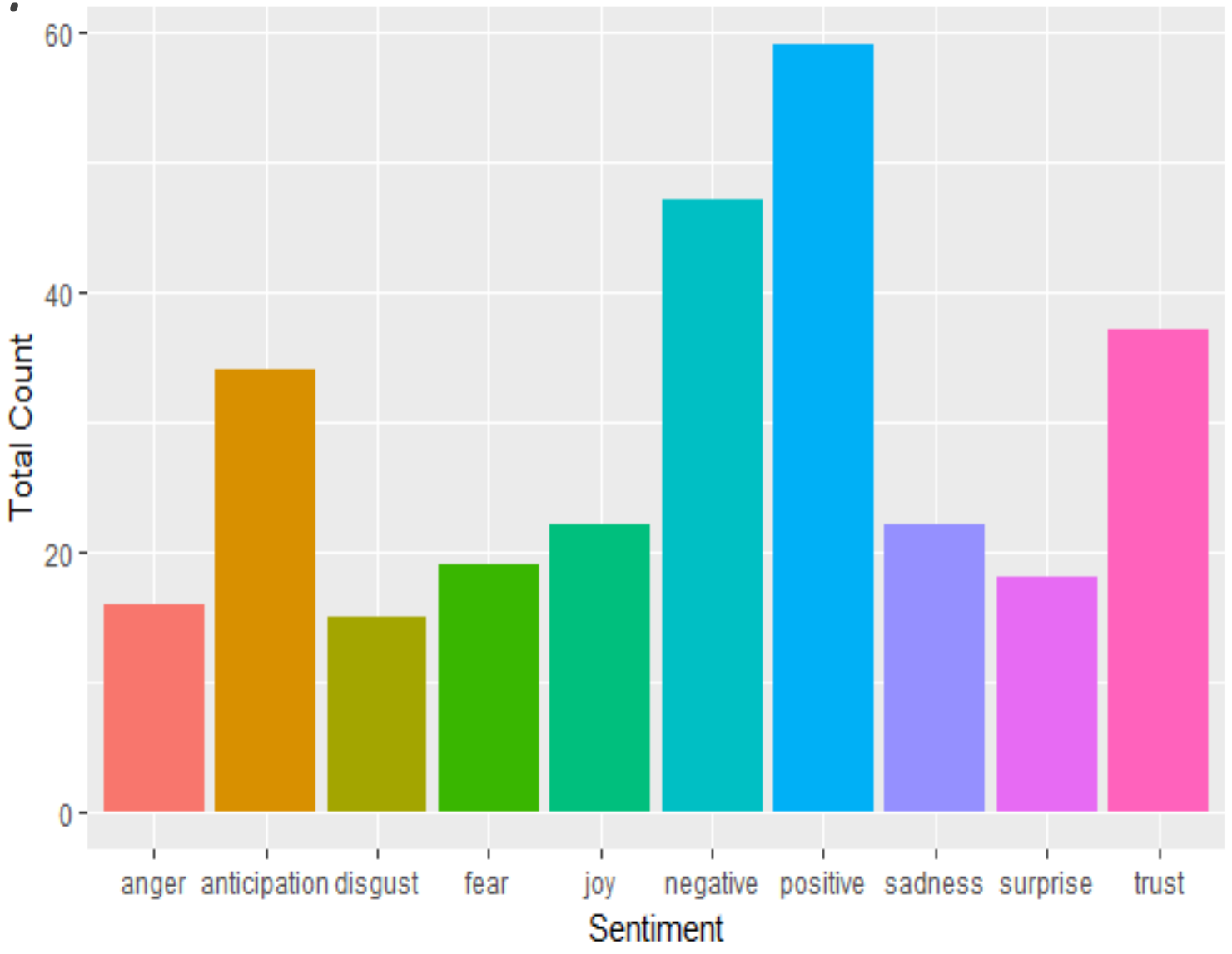
(Lui & Hu, 2005)



NRC Lexicon approach: Emotion Analysis

8 basic emotions as well as positive and negative words

(Mohammad & Turney, 2013)



More lexicons

- Available lexicons
  - General Inquirer: 1915 positive words and 2291 negative words
  - LIWC (Linguistic Inquiry and Word Count): 2300 words, >70 classes
  - MPQA Subjectivity Cues Lexicon: 2718 positive words and 4912 negative words. Each word annotated for intensity (strong, weak)
  - SentiWordNet: All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
    - [estimable(J,3)] “may be computed or estimated”: Pos=0, Neg=0, Obj=1
    - [estimable(J,1)] “deserving of respect or high regard”: Pos=0.75, Neg=0, Obj=0.25

- You can create yours – NLP Research (IgboSentilex)

Advanced (Aspect based)

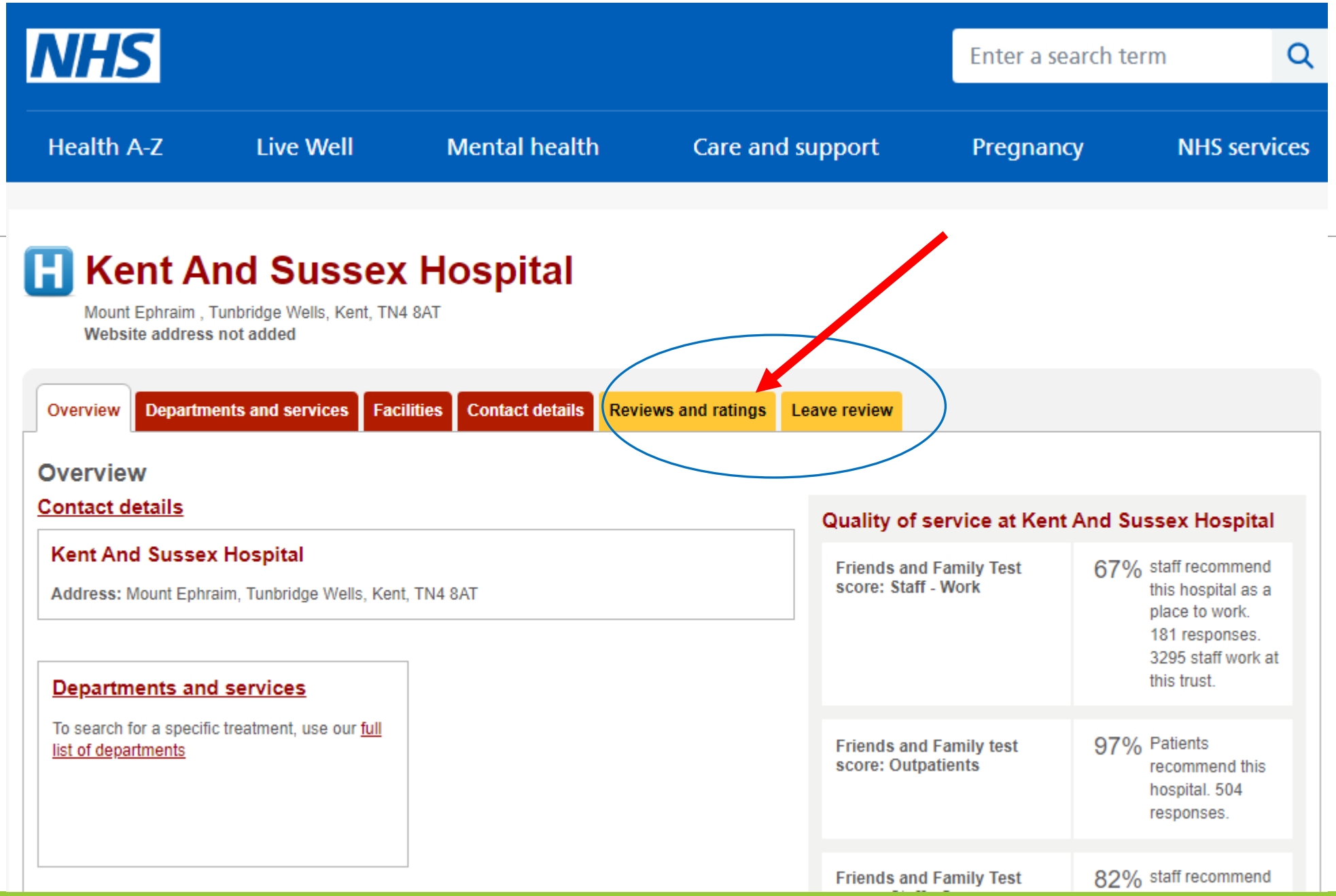
- Perfume review in *Perfumes: the Guide*: “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
  - Negative review but very difficult to figure out using just positive or negative words.

- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”
  - Seems like a positive review but it is not.

- Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is **not so good** either, I was surprised.
  - Ordering effect
  - Multiple sentiments within same sentence.

Application in Health & Security

Developed nations vs Nigeria



https://www.nhs.uk/

FEDERAL MINISTRY OF HEALTH

HOME | ABOUT US | DEPTS / UNITS | PROGRAMMES | MEDIA | RESOURCES | PARASTATALS | CONTACT

- Teaching Hospitals
- Federal Medical Centres
- Specialty Hospitals

Home > PARASTATALS > Teaching Hospitals

TEACHING HOSPITALS

COVID-19

No Reviews or rating option

Print Email

List of all Federal Teaching Hospitals:

S/N	Name of Hospital	Location	State
1	University of Benin Teaching Hospital	Benin City	Edo
2	University of Calabar Teaching Hospital	Calabar	Cross River
3	University of Abuja Teaching Hospital	Gwagwalada	FCT
4	University College Hospital	Ibadan	Oyo

- On the 6th of March 2022, 17 new confirmed cases were recorded in Nigeria
- Till date, 254,657 cases have been confirmed, 249,199 cases have been discharged and 3,142 deaths have been recorded in 36 states and the Federal Capital Territory
- The 17 new cases are reported from 5 States -Imo (8), Oyo (4), FCT (2), Osun (2) and Kano (1)
- A multi-sectoral national

https://www.health.gov.ng/

uk.trustpilot.com/review/police.uk

Trustpilot | Search for another company... | Categories | Blog | Log in | For busines

Police

1.4

police.uk | Write a review

1 review | GB

6 hours ago

Corruption in the police force

Some one in south police hav e been harassing me to keep quiet about grooming gangs and possible peadofile rings...
If you cant trust the police who can you trust .keep our kids safe from mentally disturbed animals
Dont bother contacting Sarah's law
Becaus they will hide information
Who is protecting peadofiles in south wales police ?

Date of experience: 24 September 2022

Useful | Share

GJ

Gary Jones
9 reviews | GB

Claim your profile to access Trustpilot's free business tools and connect with customers.

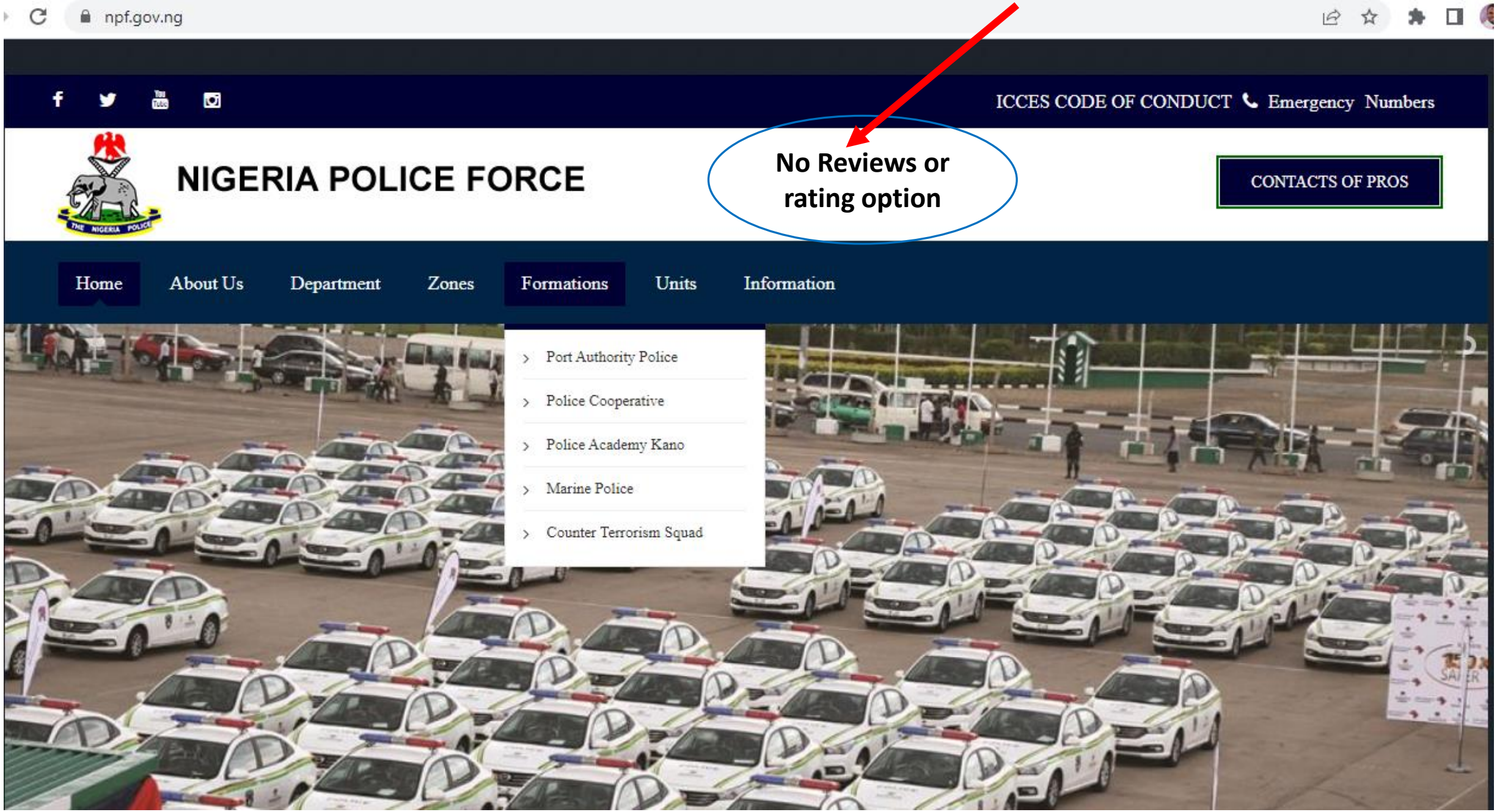
Get free account

People who looked at this company also looked at

Metropolitan Police
174
Unclaimed

Independent Office for Police Conduct
65
Unclaimed

ITV
719


npf.gov.ng
ICCES CODE OF CONDUCT
Emergency Numbers
NIGERIA POLICE FORCE
No Reviews or rating option
CONTACTS OF PROS
Home
About Us
Department
Zones
Formations
Units
Information
Port Authority Police
Police Cooperative
Police Academy Kano
Marine Police
Counter Terrorism Squad

Can Servicom play this role?

https://servicom.gov.ng/

- Statistics and Biostatistics
- Privacy
- Imaging Sciences
- Mathematical Methodologies
- Numerica
- Informati
- Modelling
- Advanced Processor Technologies
- Network Analysis

Can Nigeria take an example from this?

National Centre for Text Mining (NaCTeM)

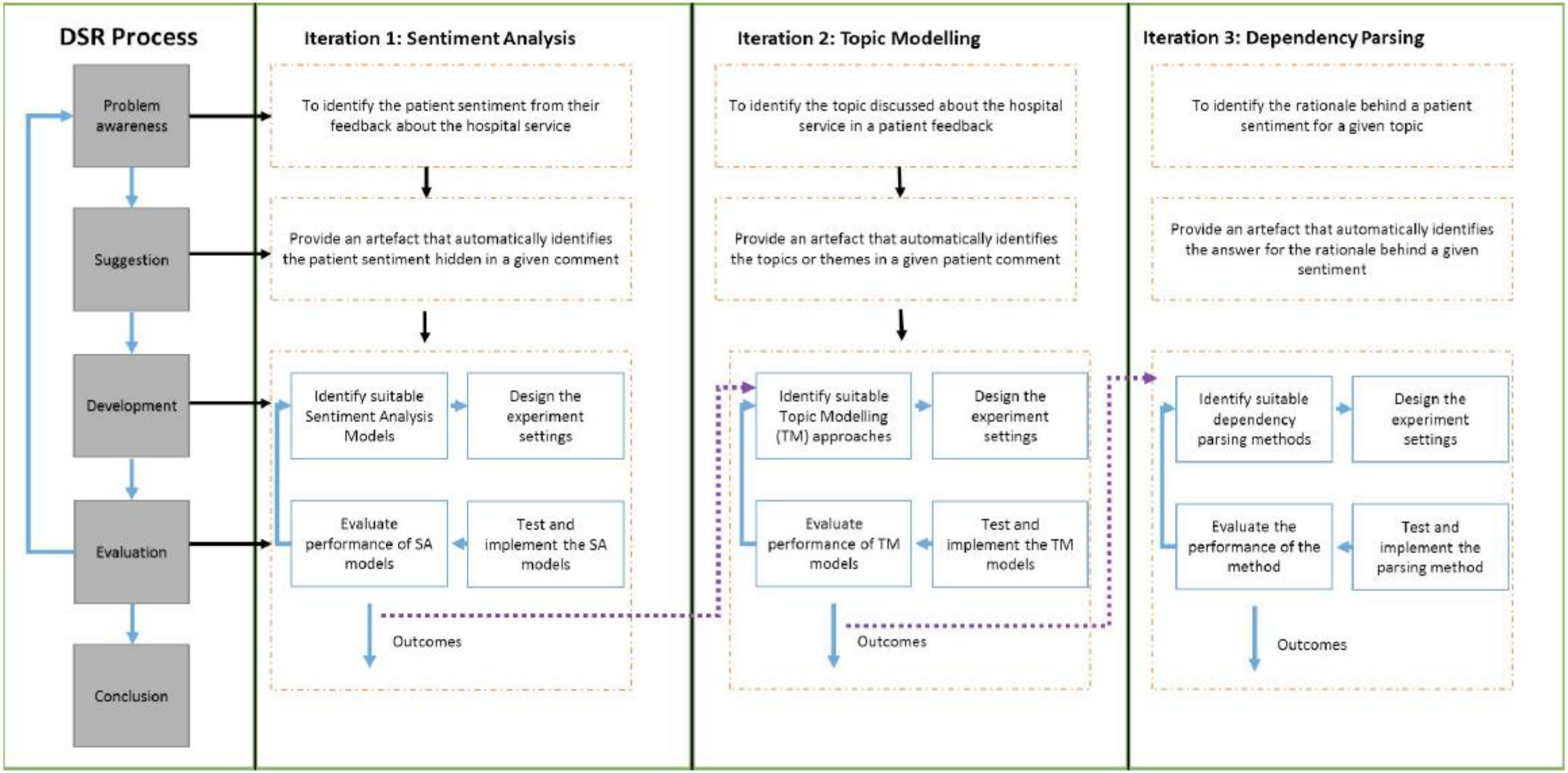
The National Centre for Text Mining (NaCTeM) is the first publicly-funded text mining centre in the world. We provide text mining services in response to the requirements of the UK academic community. Text mining offers a solution to the challenge of 'data deluge', information overload and information overlook.

NaCTeM has developed text mining services and service exemplars for the UK academic community. Our services are underpinned by a number of generic natural language processing tools:

- TerMine is a Term Management System which identifies key phrases in text.
- RobotAnalyst is a tool to minimise the human workload involved in the study identification phase of systematic reviews.
- Thalia is a semantic search engine for Pubmed abstracts.
- AcroMine is an acronym dictionary which can be used to find distinct expanded forms of acronyms from MEDLINE.
- Kleio is an advanced information retrieval system providing knowledge enriched searching for biomedicine.
- FACTA+ is a MEDLINE search engine for finding associations between biomedical concepts.
- History of Medicine (HOM) – A semantic search system over historical medical archives
- APLenty – An annotation tool for creating high-quality sequence labelling datasets using active and proactive learning
- Paladin – A document classification annotation web application which supports active/proactive learning.
- MEDIE uses semantic search to retrieve biomedical correlations from MEDLINE.

https://www.idsai.manchester.ac.uk/research/methodologies/natural-language-processing-nlp-text-mining/

Text Mining Process Flow



Source: Badja & Razaak, 2018

Sentiment Analysis Framework

1 – Data Assemble

Different Data Sources

Direct Communication
What customers are saying!

Call Center

Online Chat

Inferred Behavior
What actions customers are taking and where!

Feedback

Email & SMS

Web & Mobile

Social Communication
What customers are saying socially

Social Media

Corpus

2 – Data Preprocessing

3A - Convert text to lowercase

3B – Remove

a) Numbers (if required)

b) Punctuations/Special characters

c) English stop words

d) Own stop words (if required)

e) Strip whitespace

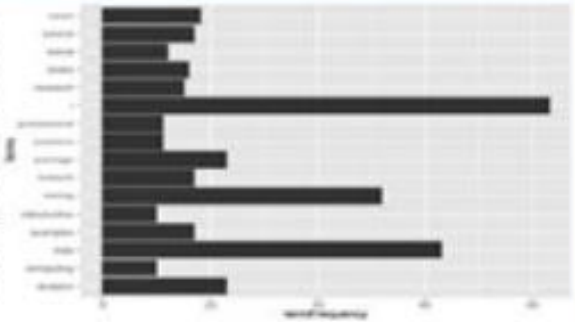
f) Stemming / Lemmatization

g) Sparse terms

2C - Create document term matrix

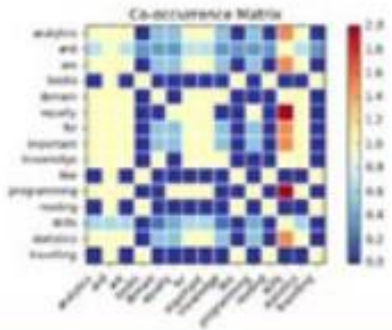
3 – Data Exploration

Frequent Terms



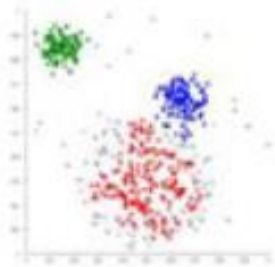
Word Cloud

Co-occurrence Plot

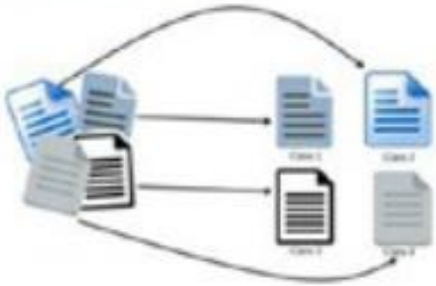


4 – Build Model

Clustering/Topic Model



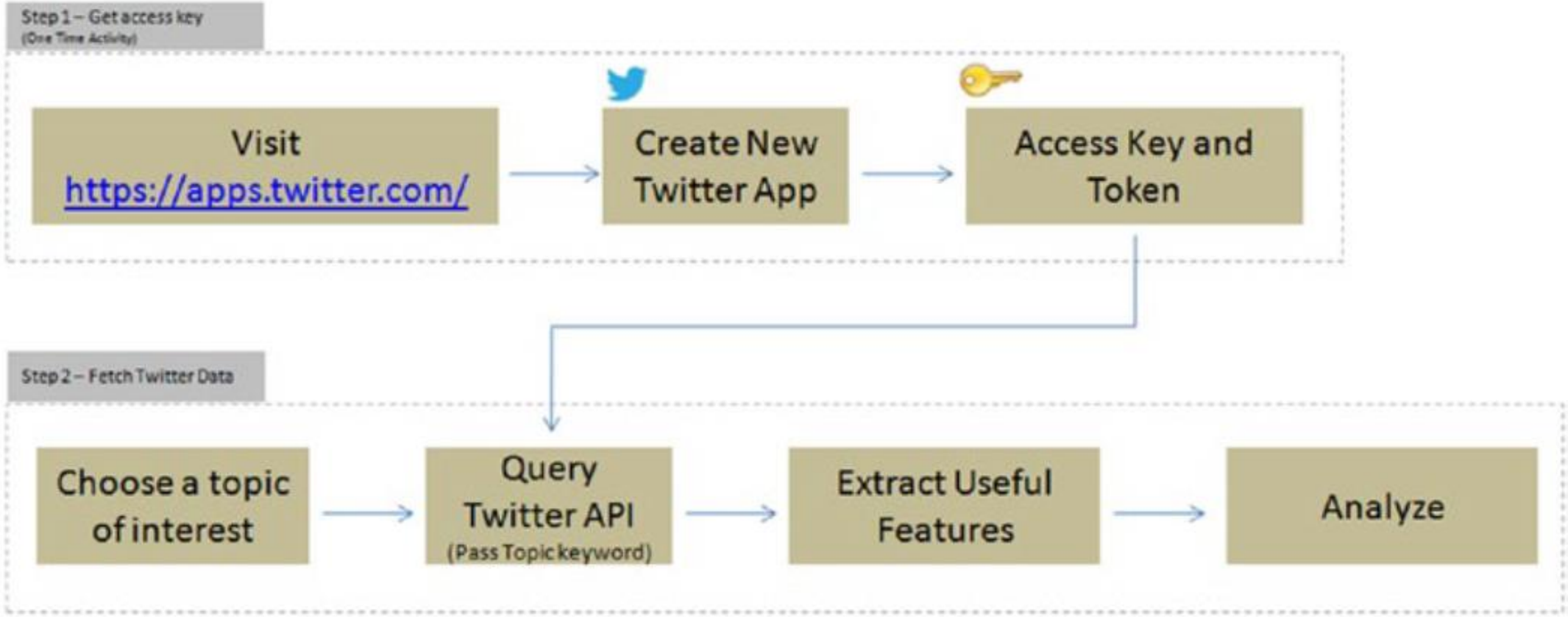
Classification



Sentiment Analysis

Source: Swamynathan, 2017

Data Assemble



Source: Swamynathan, 2017

Preprocessing

1. Tokenizing is the process of breaking a large set of texts into smaller meaningful chunks such as sentences, words, and phrases.

2. Remove all information (noise) that is not comparative or relevant to text analytics. Most common noises are numbers, punctuations, stop words, white space, etc

3. PoS tagging is the process of assigning language-specific parts of speech such as nouns, verbs, adjectives, and adverbs, etc., for each word in the given text.

4. Stemming is the process of transforming to the root word, that is, it uses an algorithm that removes common word endings from English words, such as “ly,” “es,” “ed,” and “s.”

5. Lemmatization is the process of transforming to the dictionary base form. For this you can use WordNet, which is a large lexical database for English words that are linked together by their semantic relationships.

Source: Swamynathan, 2017

Term frequency will tell you how frequently a given term appears.

$$\mathrm{TF}\ (\mathrm{term}) = \frac{Number\ of\ times\ term\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

For example, consider a document containing 100 words wherein the word 'ML' appears 3 times, then TF (ML) = 3 / 100 =0.03

Document frequency will tell you how important a term is?

$$\mathrm{DF}\ (\mathrm{term}) = \frac{d\,(number\ of\ documents\ containing\ a\ given\ term)}{D\,(the\ size\ of\ the\ collection\ of\ documents)}$$

Source: Swamynathan, 2017

Assume we have 10 million documents and the word ML appears in one thousand of these, then DF (ML) = 1000/10,000,000 = 0.0001

To normalize let's take a log (d/D), that is, log (0.0001) = -4

Quite often D > d and log (d/D) will give a negative value as seen in the above example. So to solve this problem let's invert the ratio inside the log expression, which is known as Inverse document frequency (IDF). Essentially we are compressing the scale of values so that very large or very small quantities are smoothly compared.

$$\text{IDF (term)} = \log\left(\frac{\textit{Total number of documents}}{\textit{Number of documents with a given term in it}}\right)$$

Continuing with the above example, IDF(ML) = log(10,000,000 / 1,000) = 4

TF-IDF is the weight product of quantities, that is, for the above example TF-IDF (ML) = 0.03 * 4 = 0.12

Source: Swamynathan, 2017

Lab I: Example Text Mining In Health

- Federal Ministry of Health - @fmohnigeria, #FMOH
- Nigeria Centre for Disease Control and Prevention - @ncdcgov
- National Primary Healthcare Development Agency - @nphcdang

The interest here is to perform opinion mining and analyze the public sentiments of these agencies.

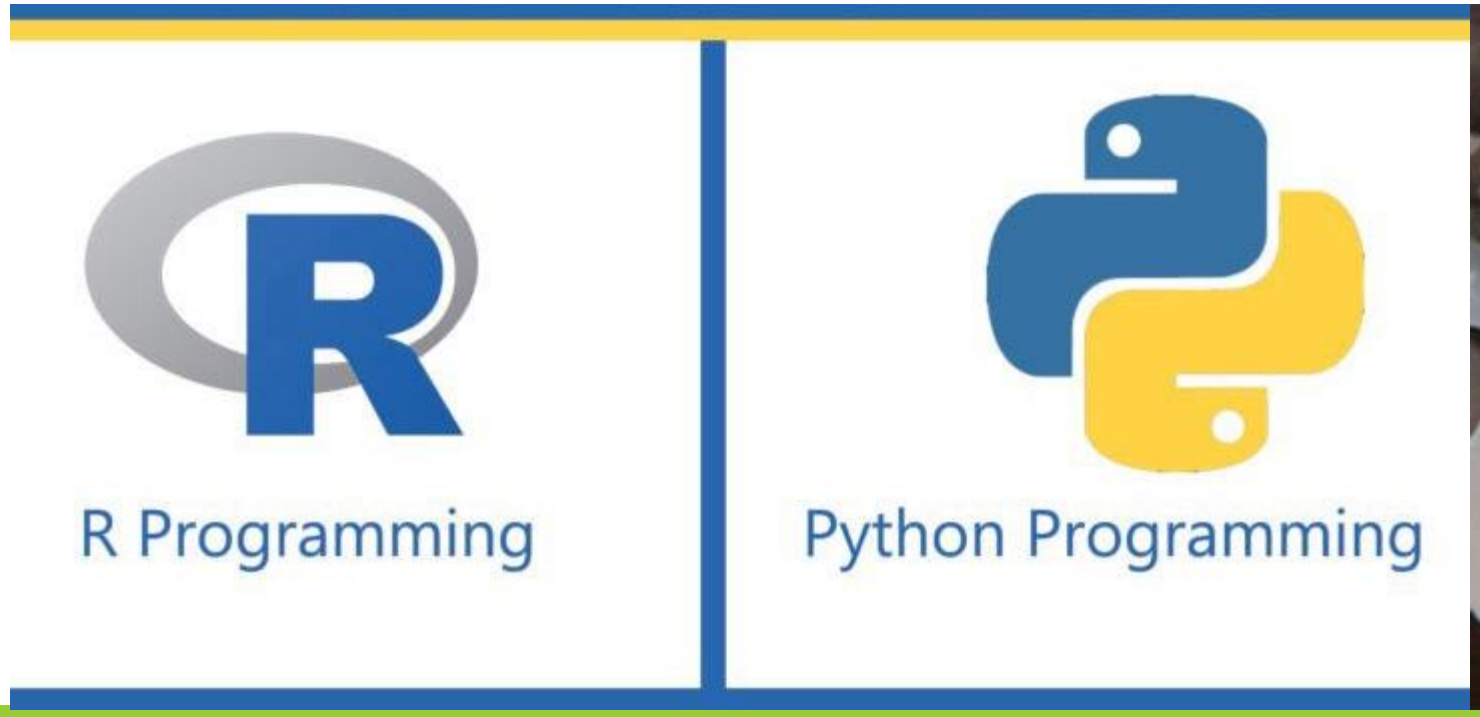
Lab II: Example Text Mining In Security

- Nigeria Police Force - @policeng, #policeng
- Nigerian Army - @hqnigerianarmy

The interest here is to perform opinion mining and analyze the public sentiments of these agencies.

Demo – 2 in R

The codes are available at the end of the lab demos



Thank you

Dr. Emeka Ogbuju

Department of Computer Science, Federal University Lokoja

Ph: 09030797424

E-mail: emeka.ogbuju@fulokoja.edu.ng