Natural Language Processing:

Learning from Kenyans on Twitter

October 8, 2021

About Me

- My ML journey started in Marketing - I am now an MSc. Social Statistics candidate studying the Kenyan informal sector
- I love learning - I enjoy being either the teacher or the student
- Life can be an exciting adventure if you make it one - I hike and cycle

Agenda

1. The problem
    - Language evolution
    - Kenya's linguistic identity hard to define
    - Current NLP approaches in Kenya not working
2. The dataset
    - Building the #KOT dataset & selecting "English" tweets
    - Exploring the #KOT dataset
3. The solution
    - Multi-model NLP
    - Multi-represented corpus
4. Benefits of using this solution

The problem

Kenyans’ linguistic identity is hard to define...

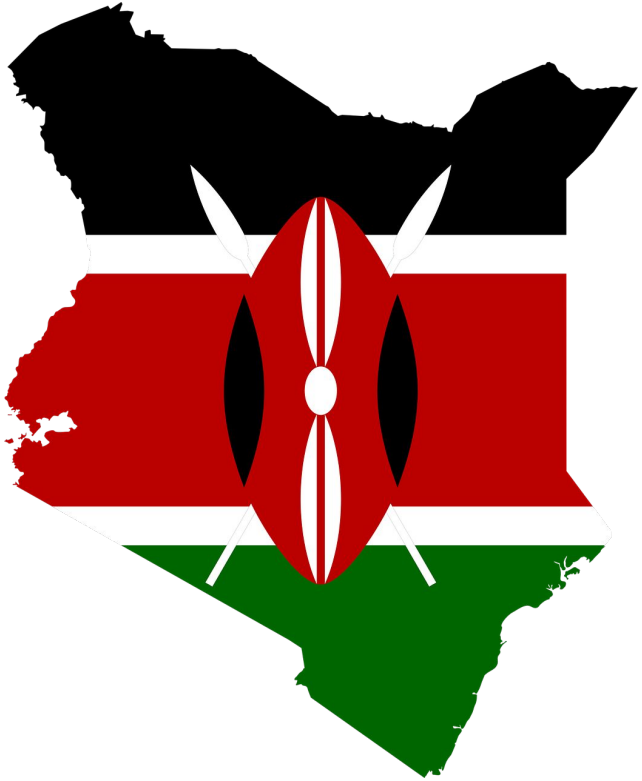
“A language family is a group of languages related through descent from a common ancestral language”

Wikipedia

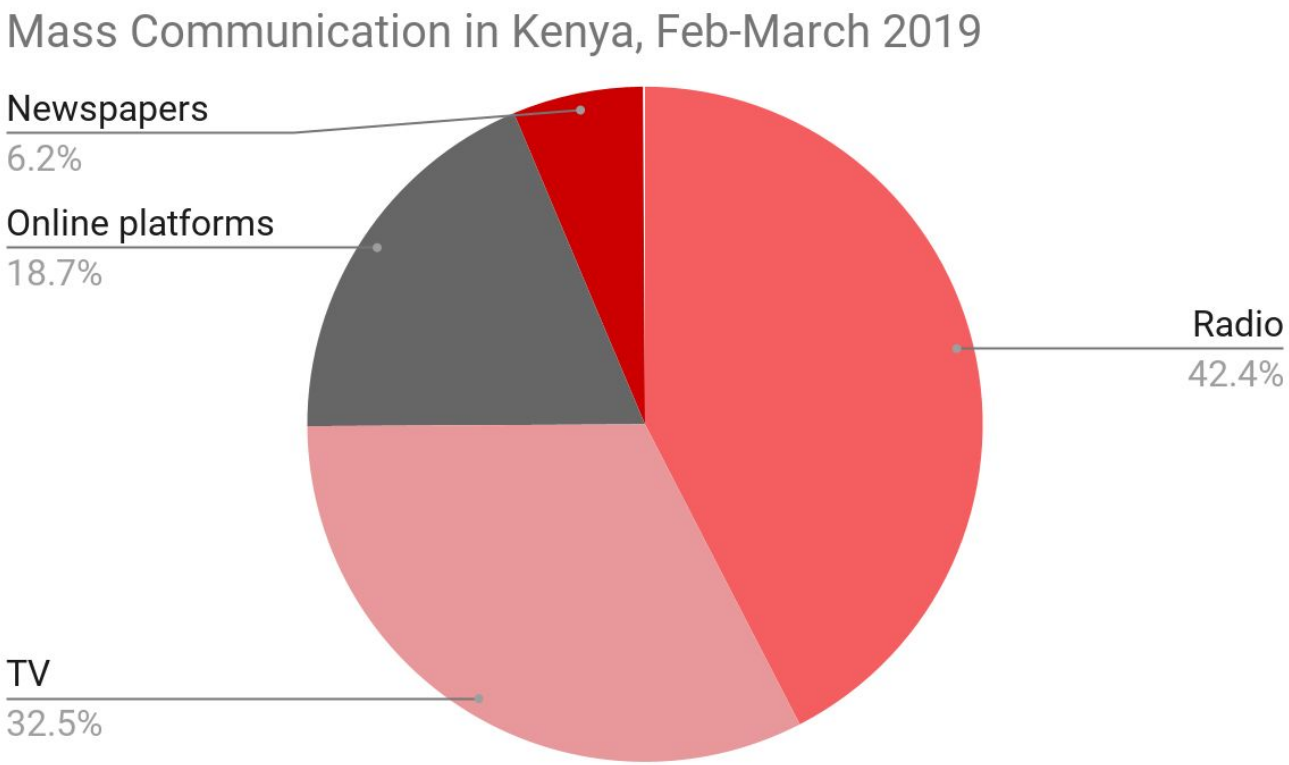
https://en.wikipedia.org/wiki/Language_family

Language families in Kenya

- Bantu, Niger-Congo branch: spoken by 65%
- Nilo-Saharan: spoken by 31%
- Afroasiatic
- Other foreign languages, especially English
- Urban dialect - Sheng



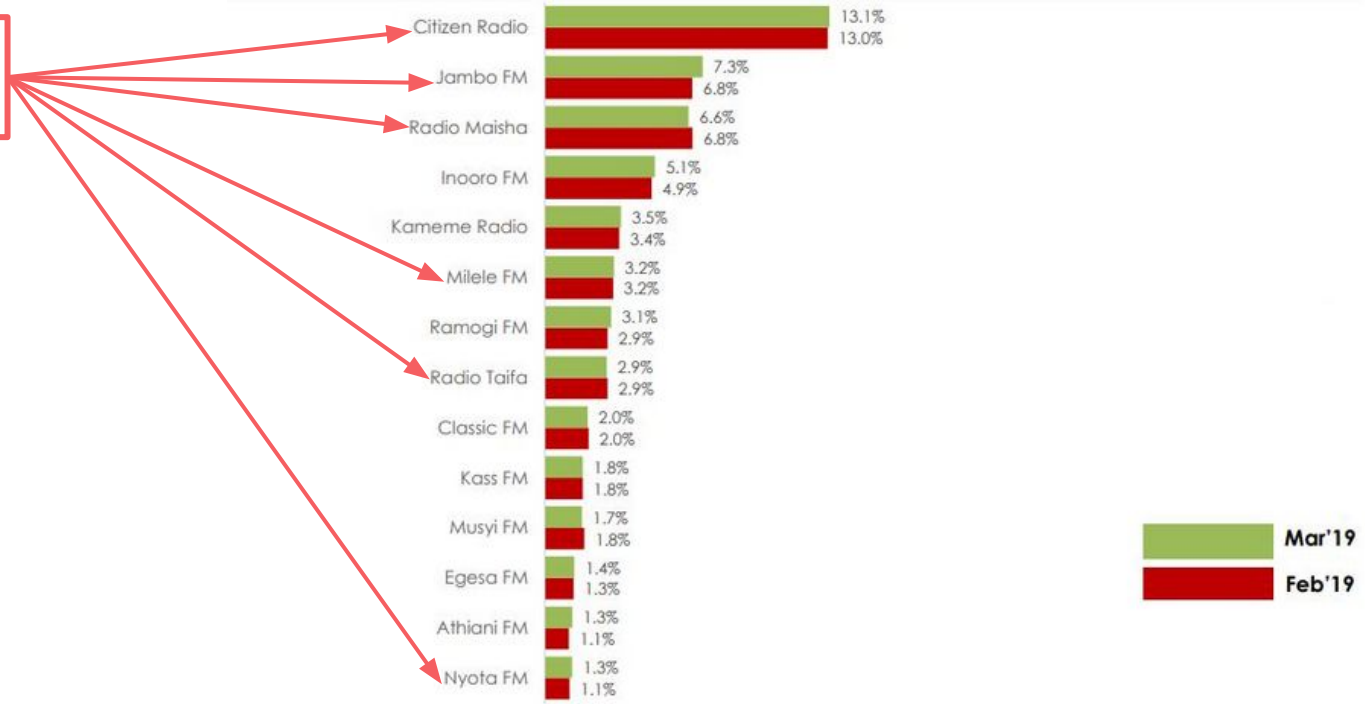
https://autolingual.com/kenya-languages/



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

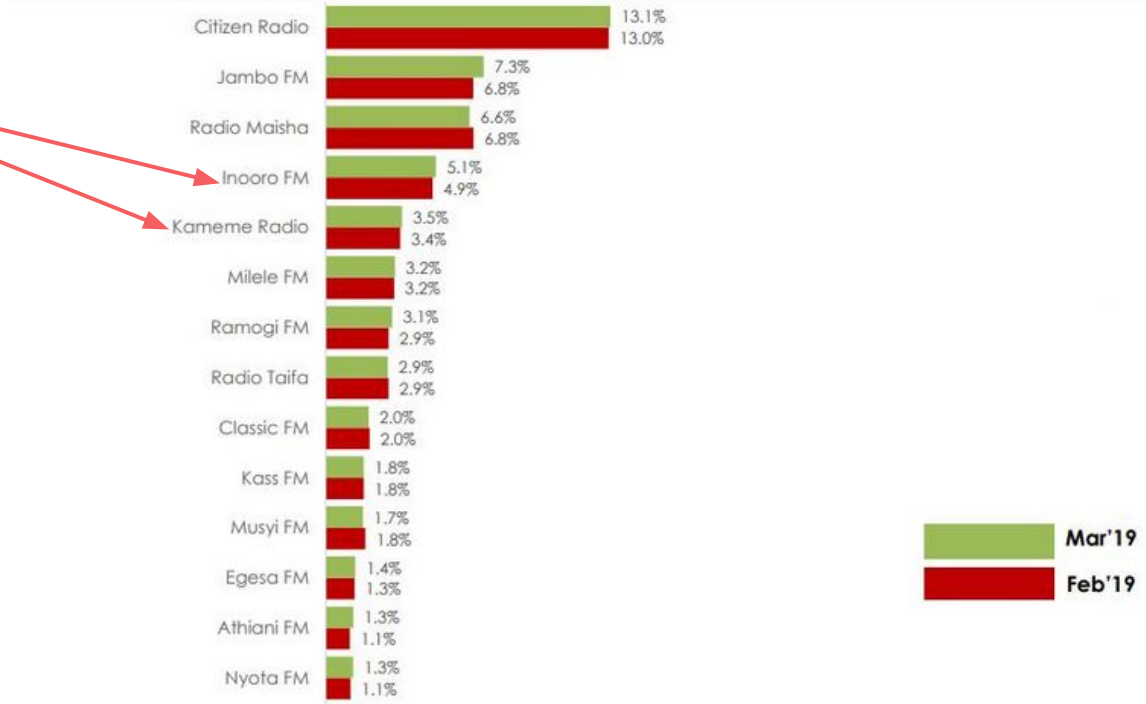
Kiswahili (Bantu)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

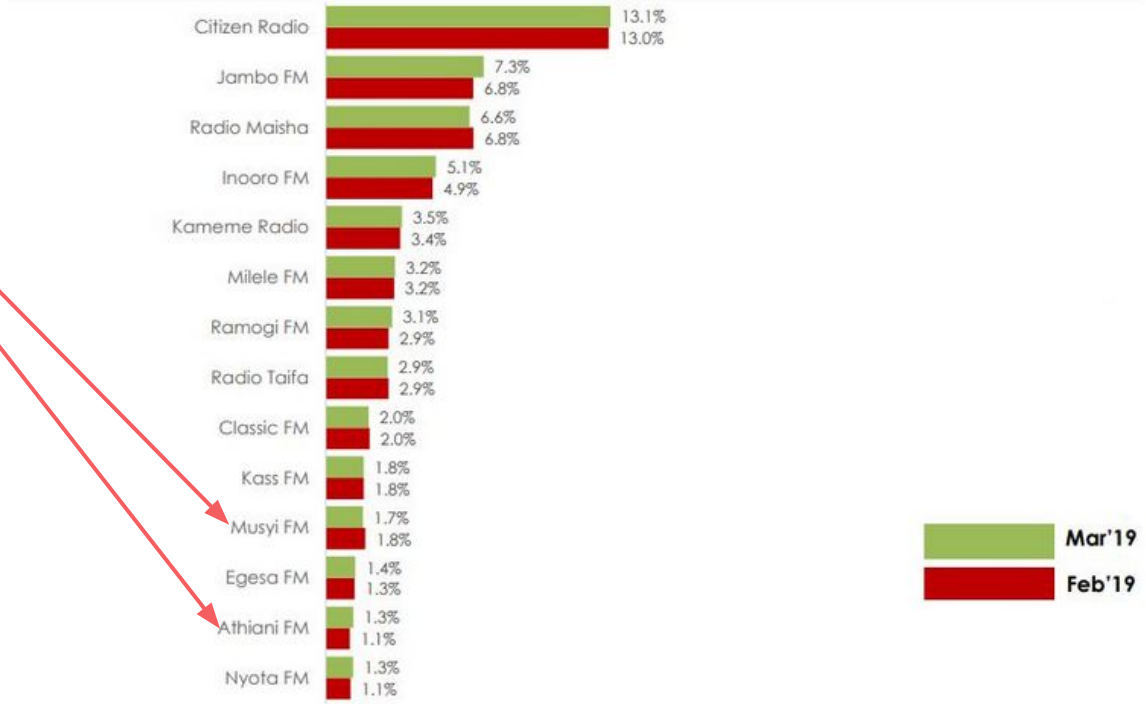
Gîkûyû (Bantu)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

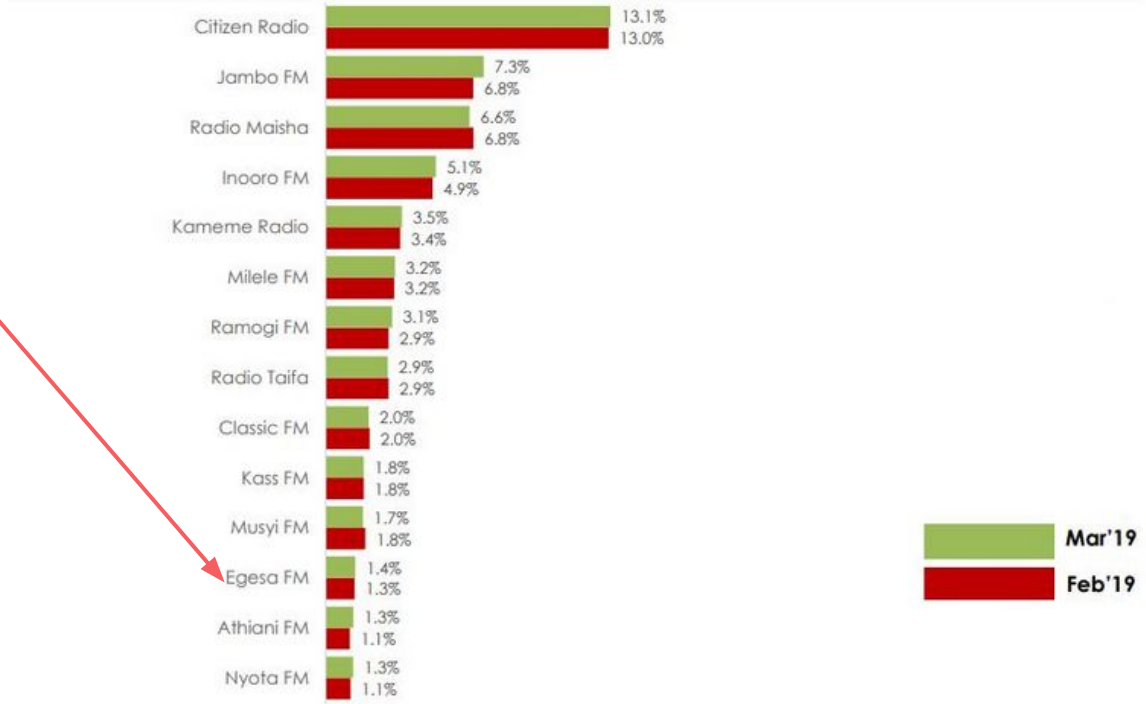
Kamba (Bantu)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

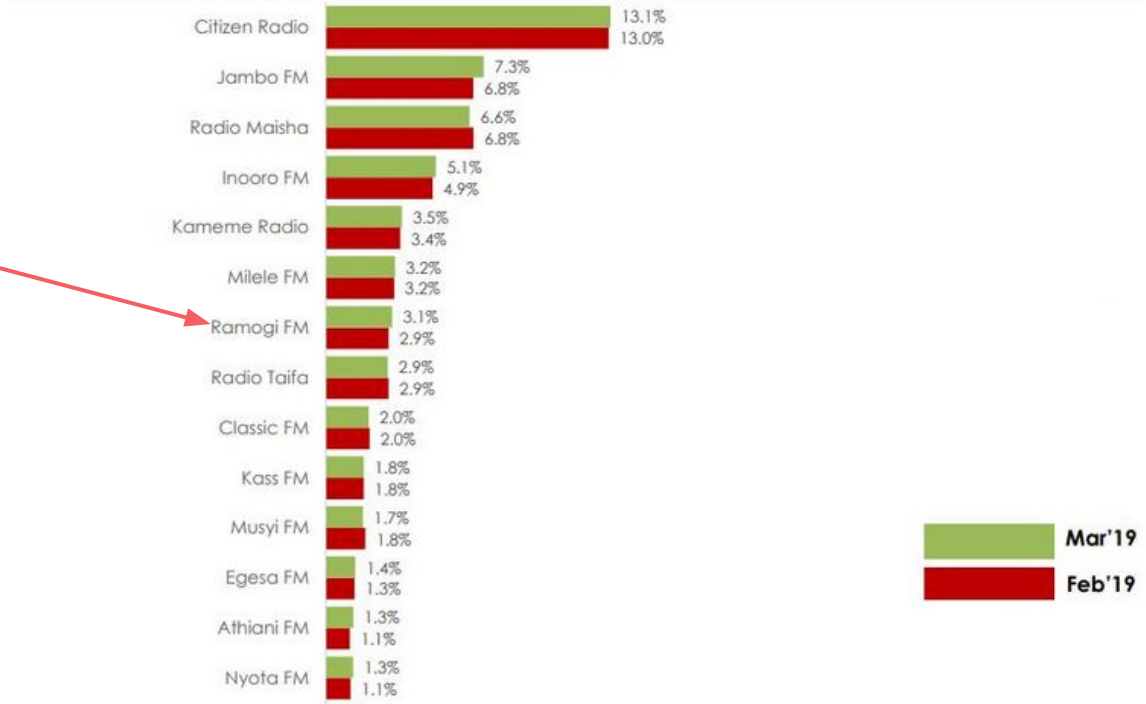
Ekegusii (Bantu)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

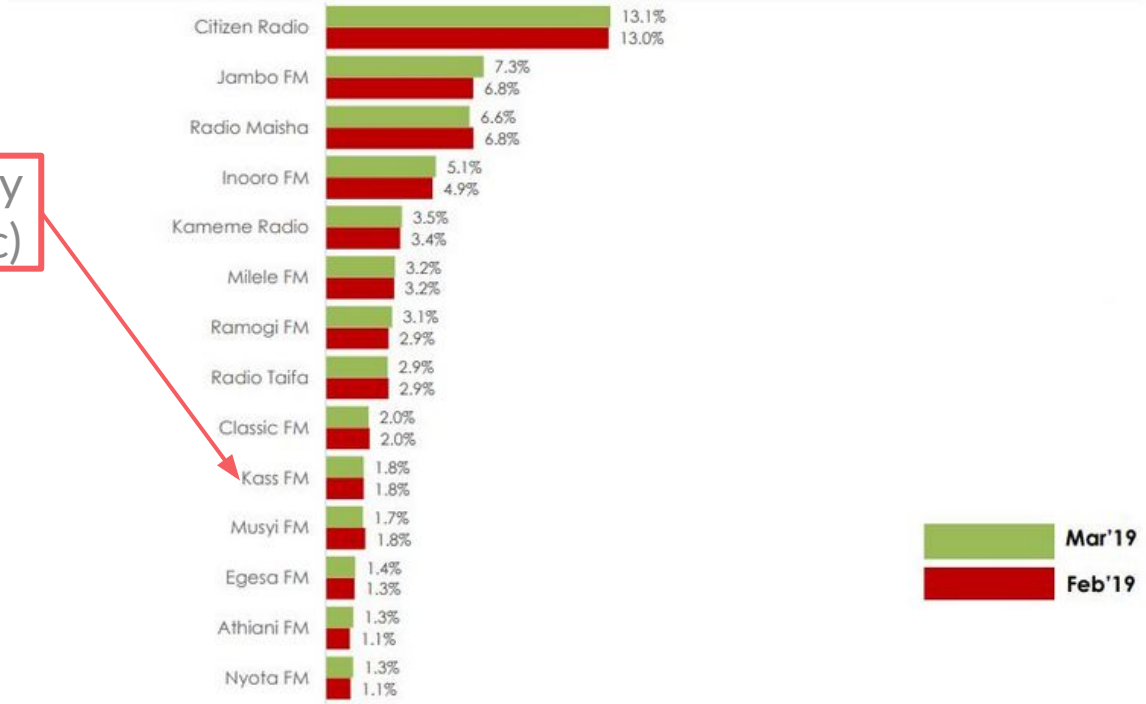
Dholuo (Nilotic)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Radio

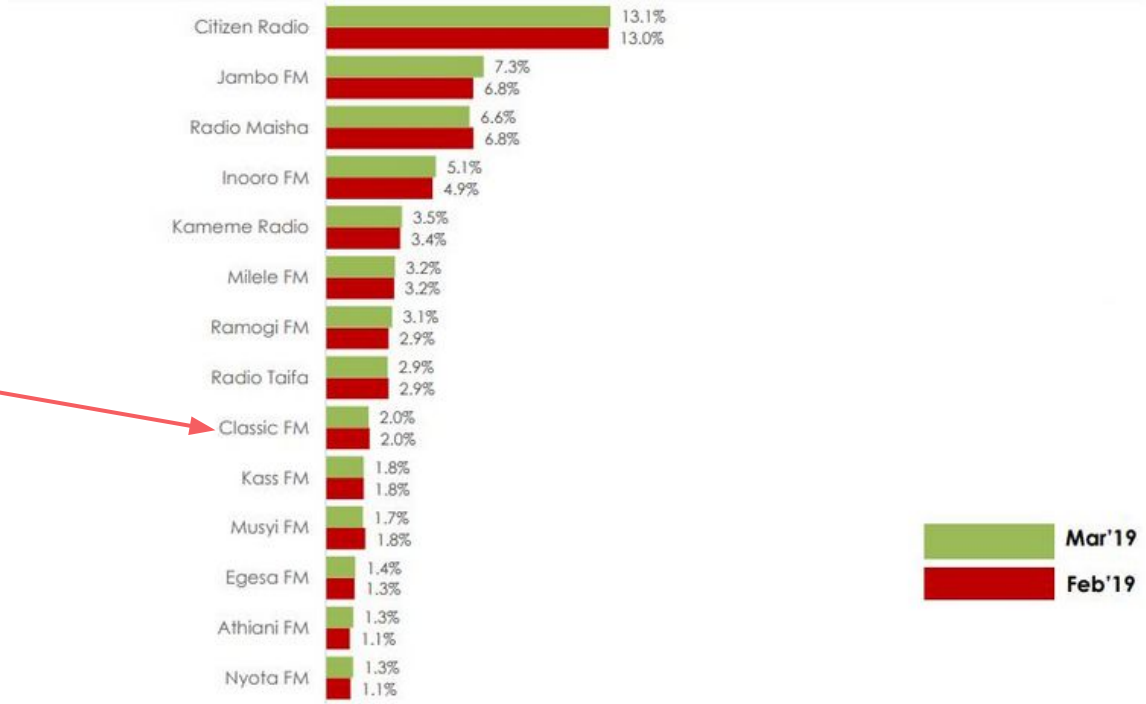
Kalenjin language family (Nilotic)



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

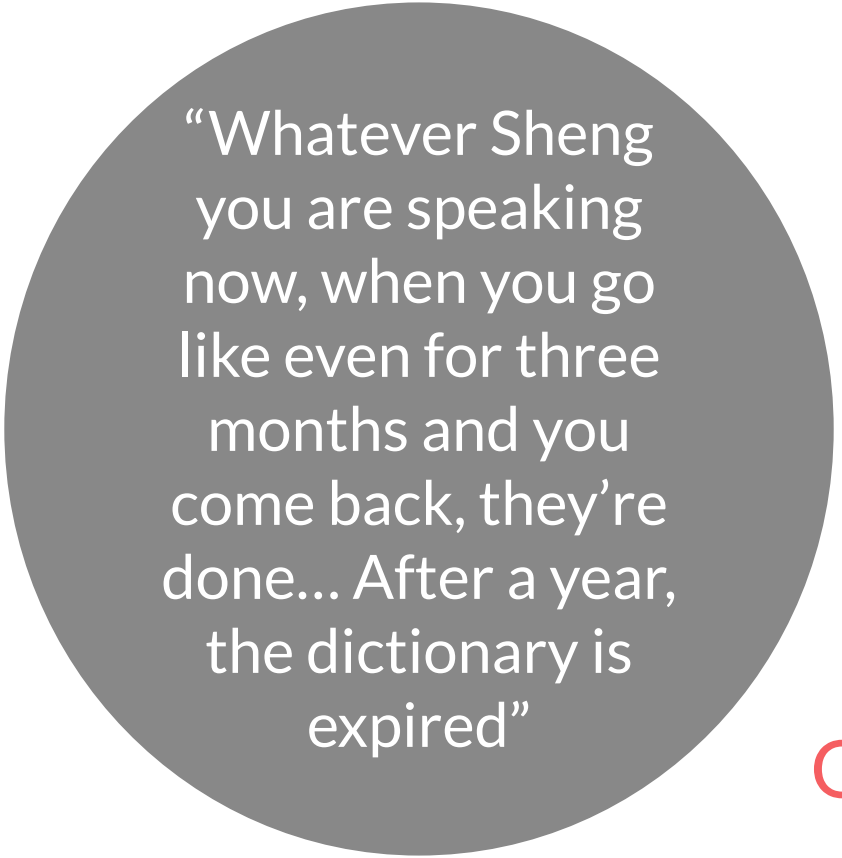
Kenyans' usage of Radio

English



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyan languages change over time, especially Sheng...



Octopizzo

https://slate.com/news-and-politics/2013/11/sheng-is-becoming-a-kenyan-language-how-the-urban-slang-of-nairobi-slums-is-spreading.html

Mamanzi

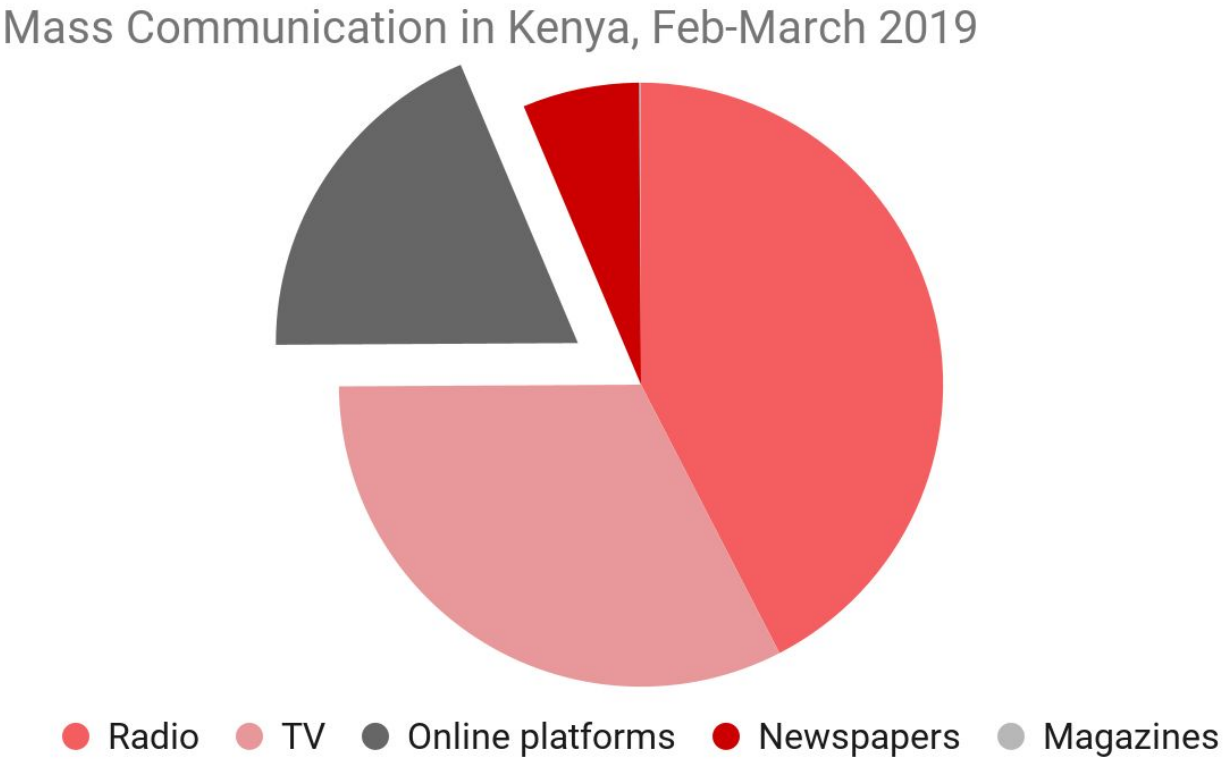
Translation: Ladies in 2001

Mayens

Translation: Ladies in 2019

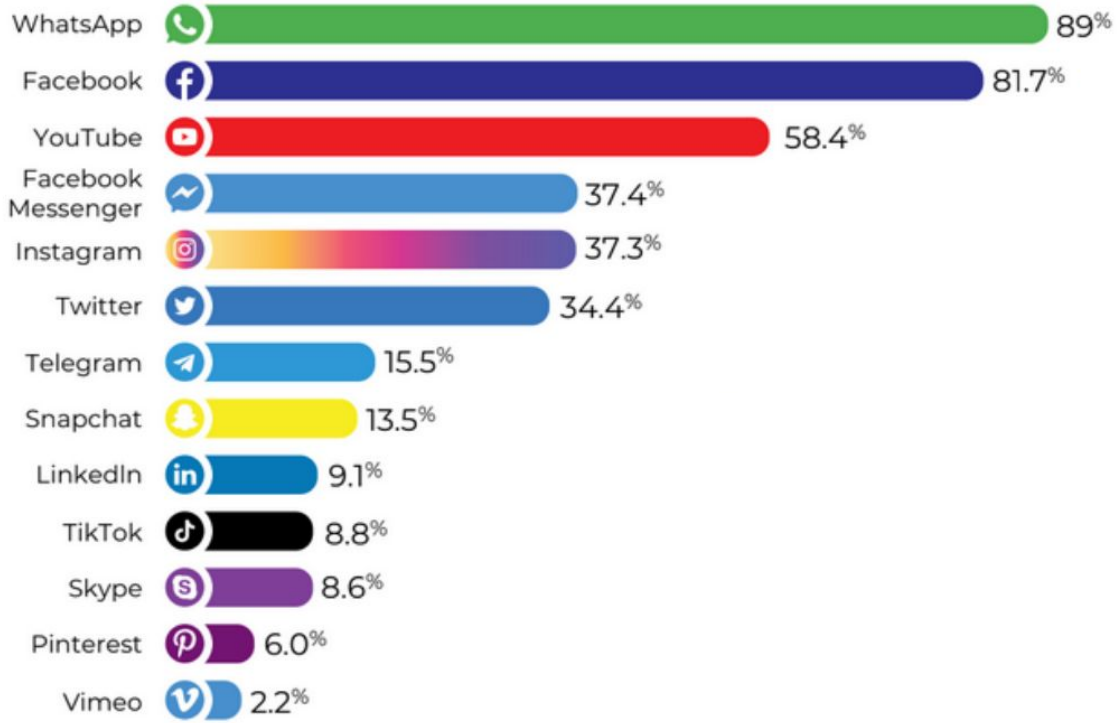
… so our approach has to be language and time agnostic.

More and more Kenyans are going online



https://businesstoday.co.ke/popular-radio-stations-citizen-radio-milele-fm/

Kenyans' usage of Social Media in 2020



https://www.usiu.ac.ke/assets/file/b8243945-usiu-africa-kenya-social-media-lansc.pdf

Kenyans' usage of Social Media

Social usage

Entertainment usage

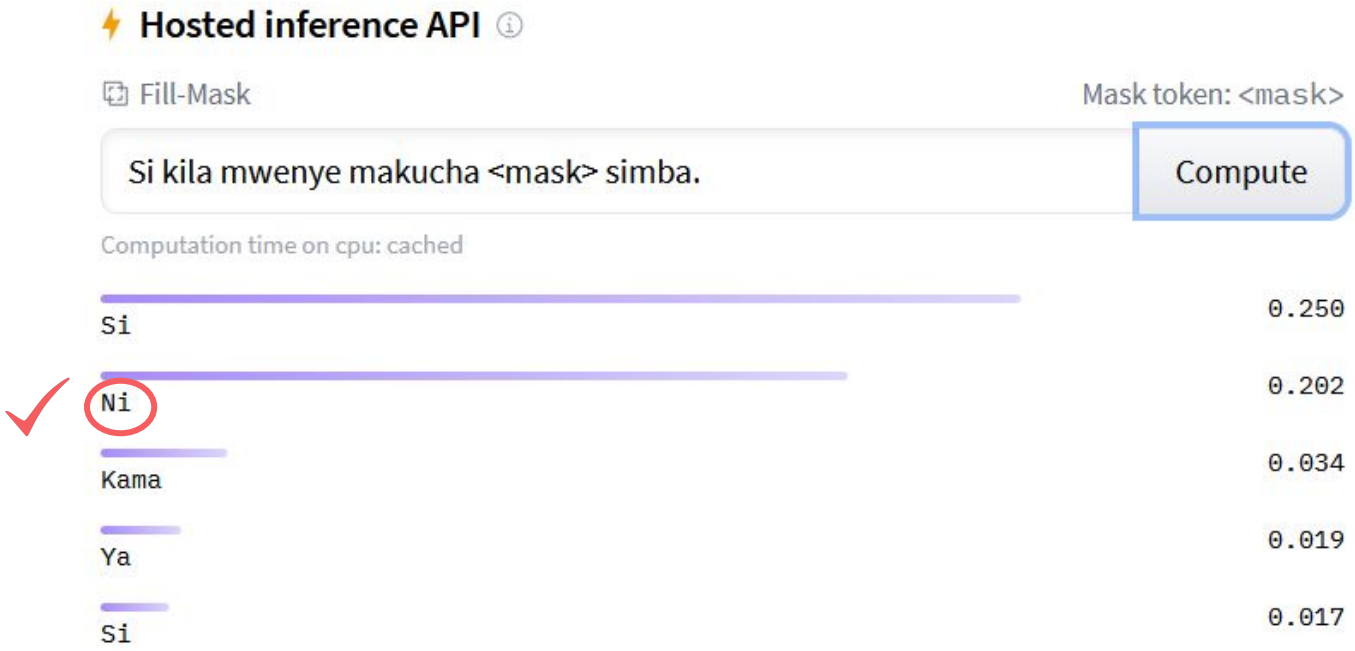
Political usage

https://www.usiu.ac.ke/assets/file/b8243945-usiu-africa-kenya-social-media-lansc.pdf

… but we don’t have good tools that can serve them.

Current NLP machine learning can’t handle Kenyan speech.

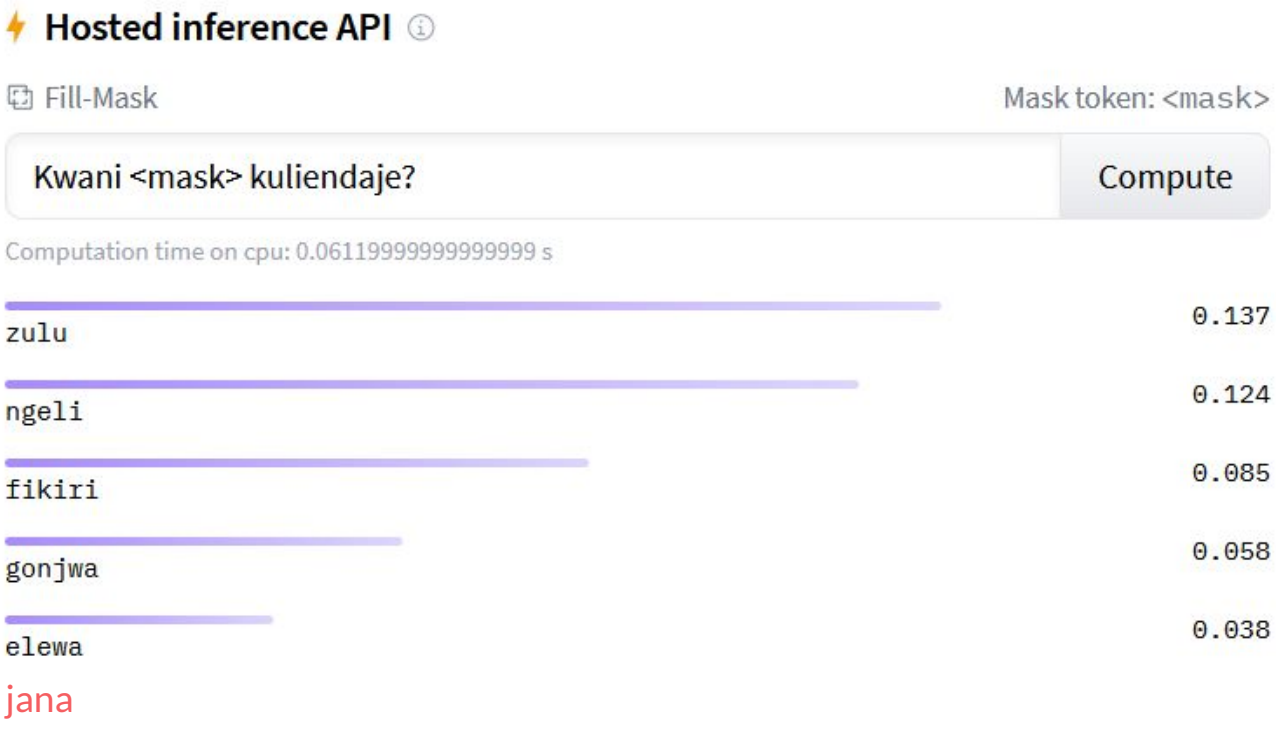
Challenges understanding Kiswahili



English Translation:
Not everyone with claws ___ a lion.
= is

https://huggingface.co/flax-community/roberta-swahili

Challenges understanding Kiswahili

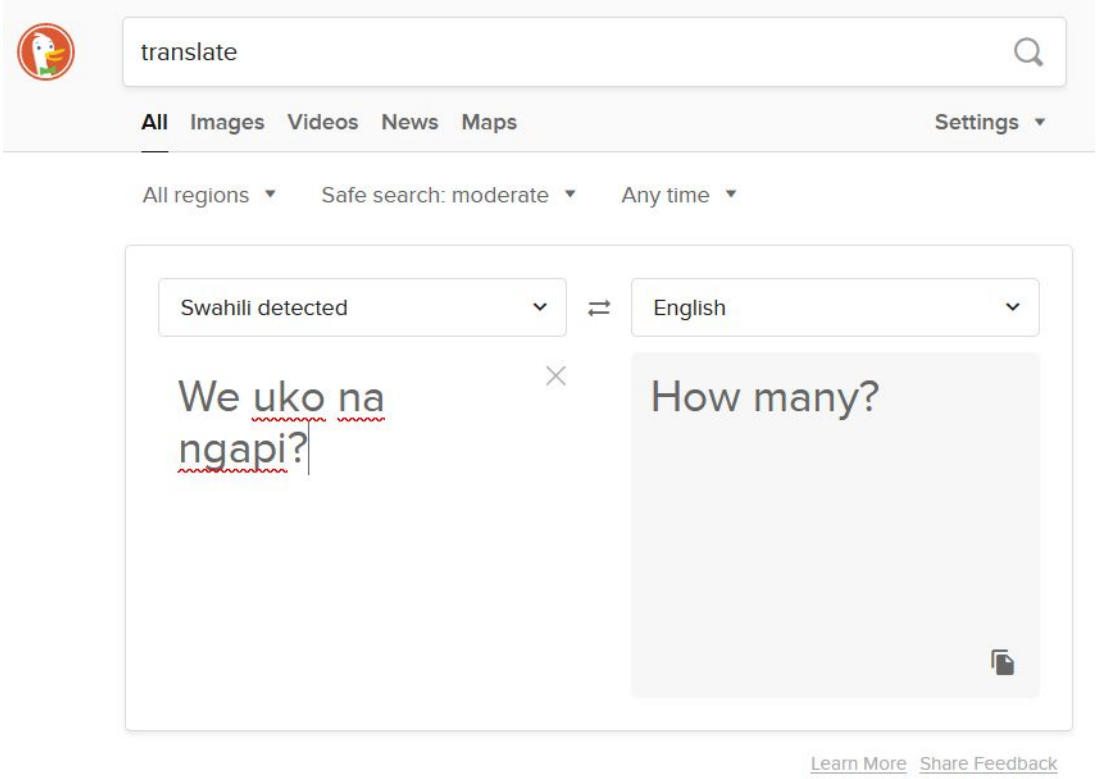


✓ jana

English Translation:
How did ___ go?
= yesterday

https://huggingface.co/flax-community/roberta-swahili

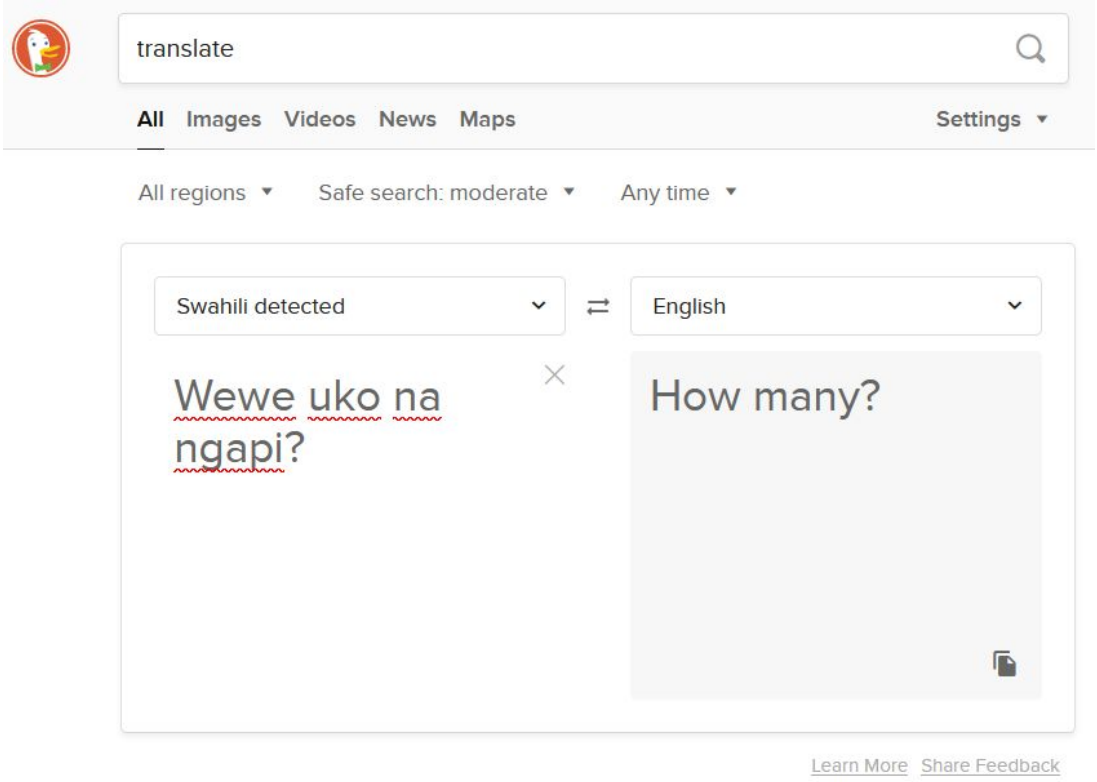
Challenges translating Kiswahili



"We" -> short form of "Wewe"

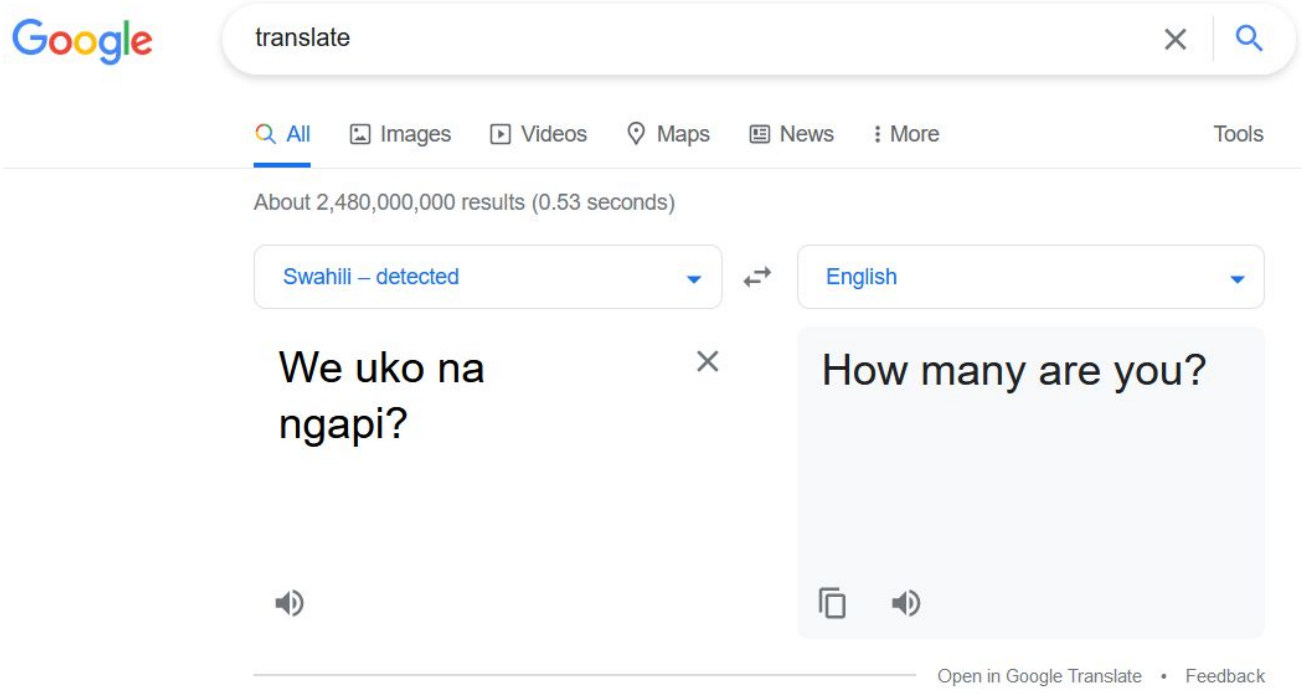
English Translation: How much/many do you have?

Challenges translating Kiswahili



English Translation: How much/many do you have?

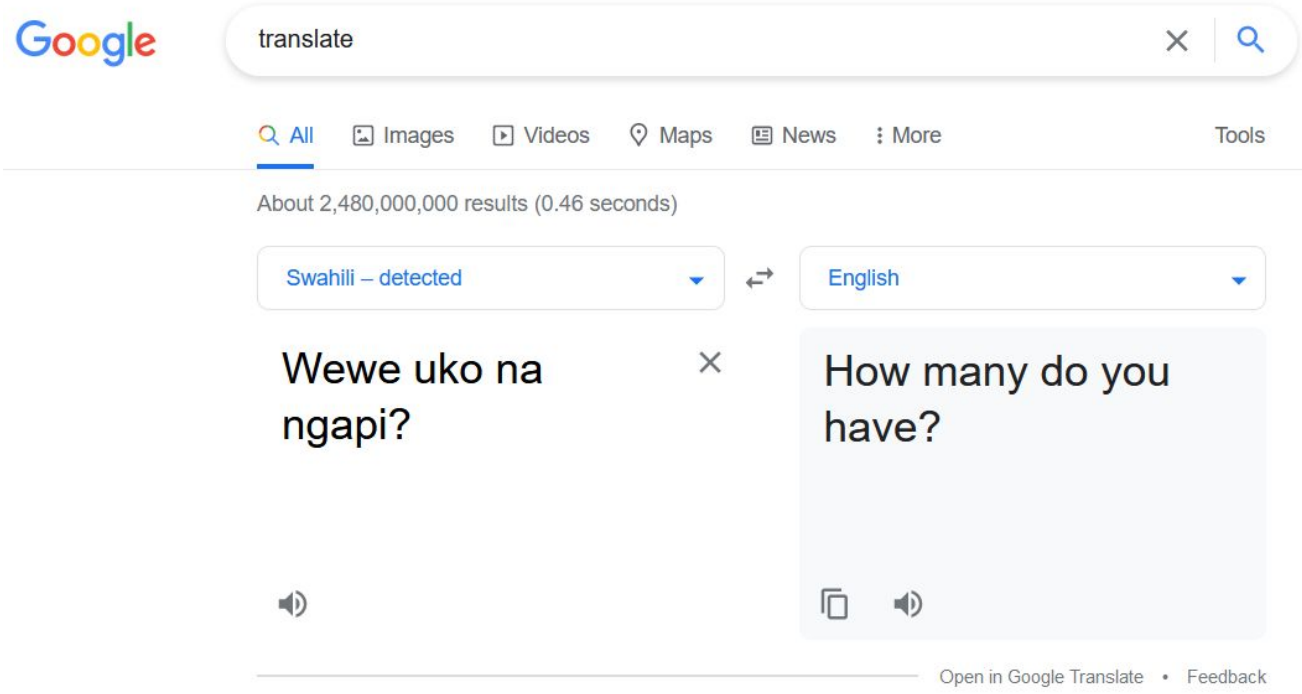
Challenges translating Kiswahili



"We" -> short form of "Wewe"

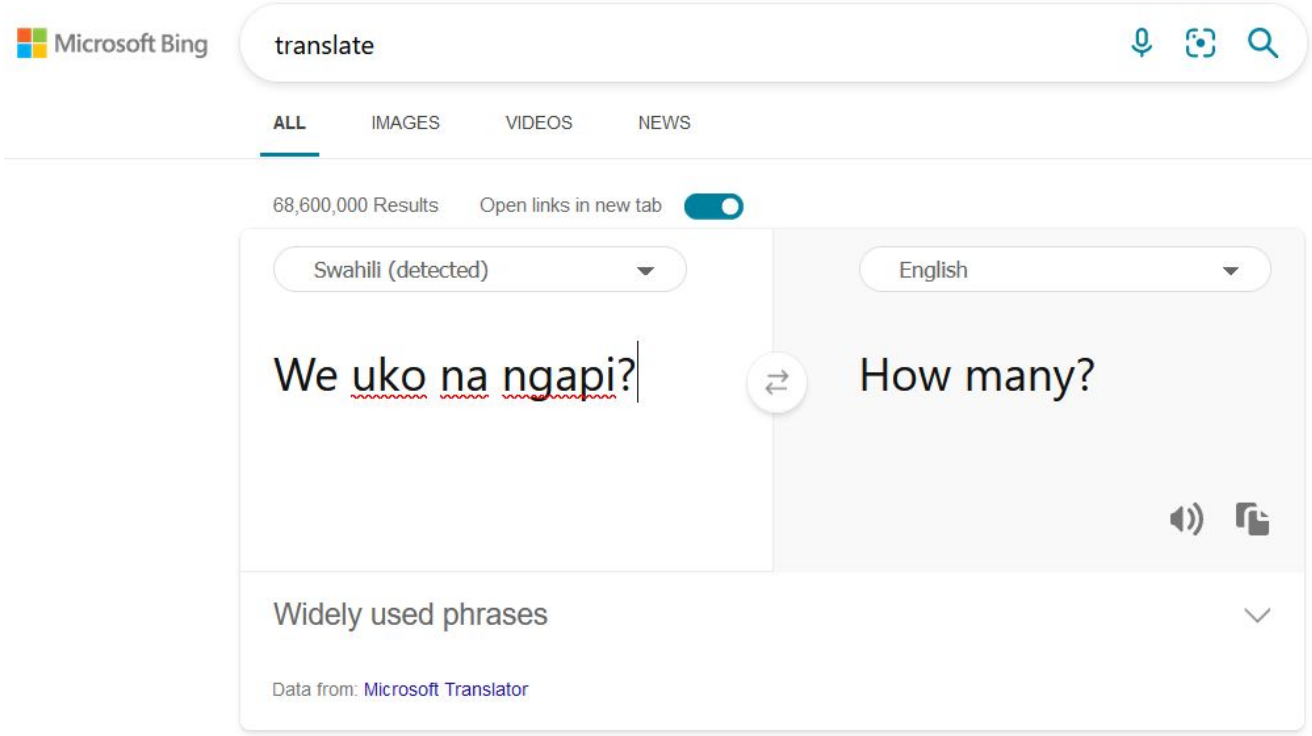
English Translation: How much/many do you have?

Challenges translating Kiswahili



English Translation: How much/many do you have?

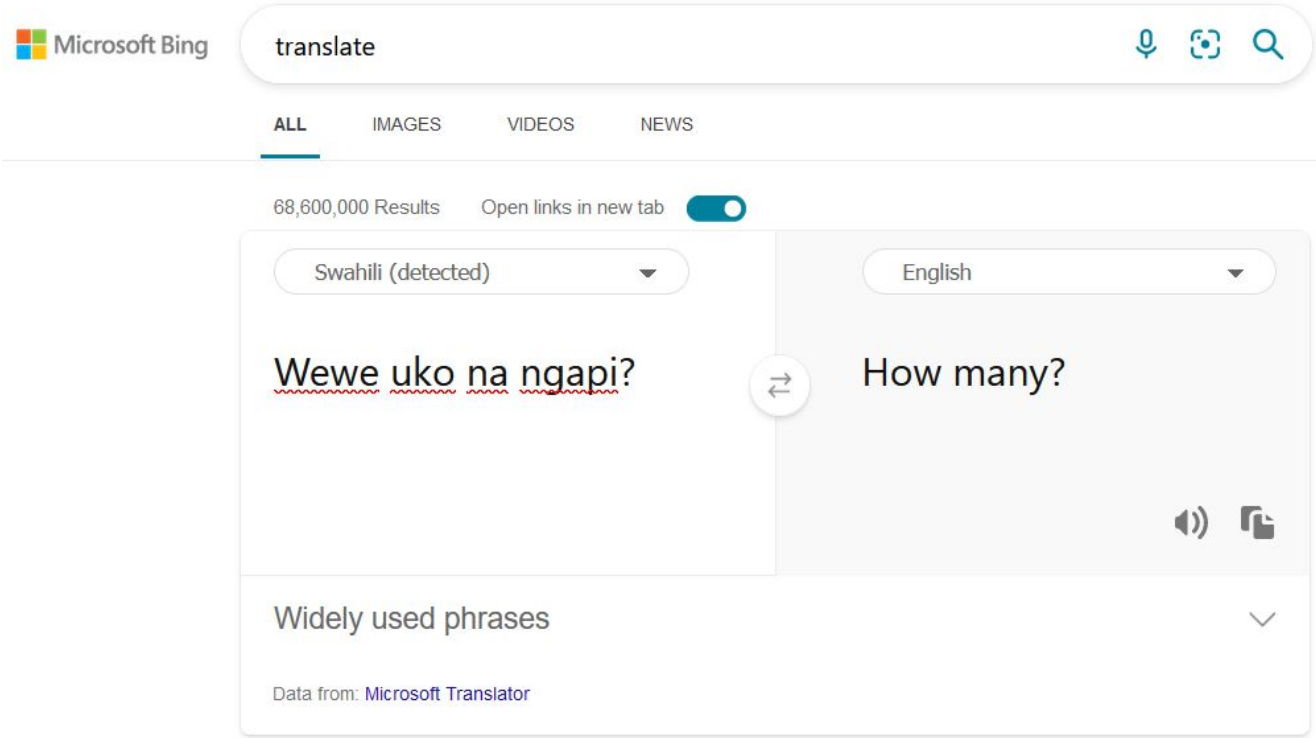
Challenges translating Kiswahili



"We" -> short form of "Wewe"

English Translation: How much/many do you have?

Challenges translating Kiswahili

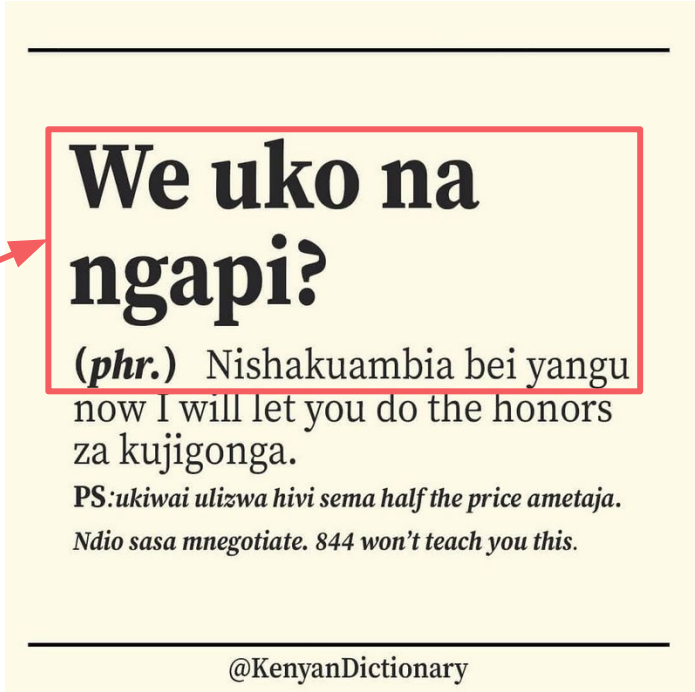


English Translation: How much/many do you have?

Challenges faced understanding Kenyans on Twitter

Code-switching
This is in the same language: Swahili

English translation:
How much do you have?
(phrase) I've already told you my price



https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

Challenges faced understanding Kenyans on Twitter

Slang

A Sheng word that means
“ to play yourself”
or
a Swahili word that means
“to hit yourself”

In this context it’s Sheng.

https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

Challenges faced understanding Kenyans on Twitter

Slang
m + negotiate

English translation:
You negotiate with each other

https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

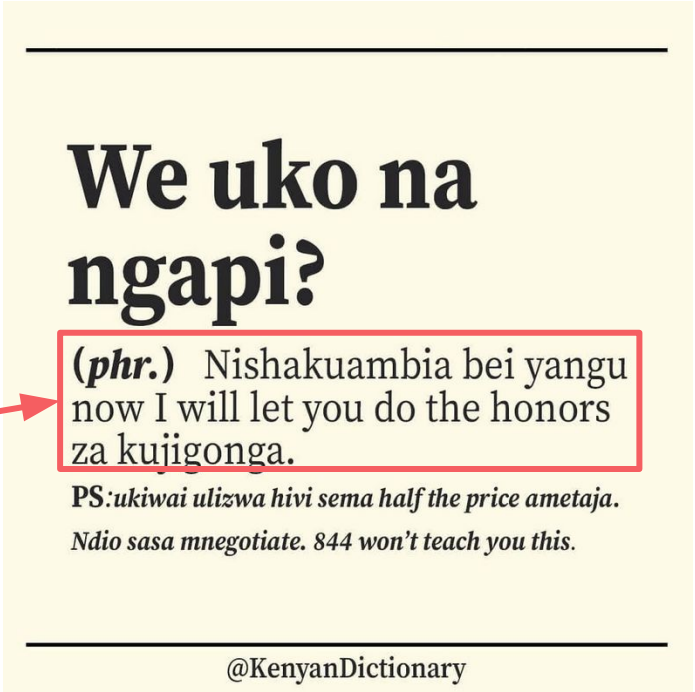
Challenges faced understanding Kenyans on Twitter

Multilingualism

Combining English, Swahili and Sheng.

English translation:

(phrase) I've already told you my price now I will let you do the honours of playing yourself.



https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

Challenges faced understanding Kenyans on Twitter

Multilingualism

Combining English, Swahili and Sheng.

English translation:

PS: if you are ever asked this question say half the price that they have stated.

So that now you negotiate with each other.

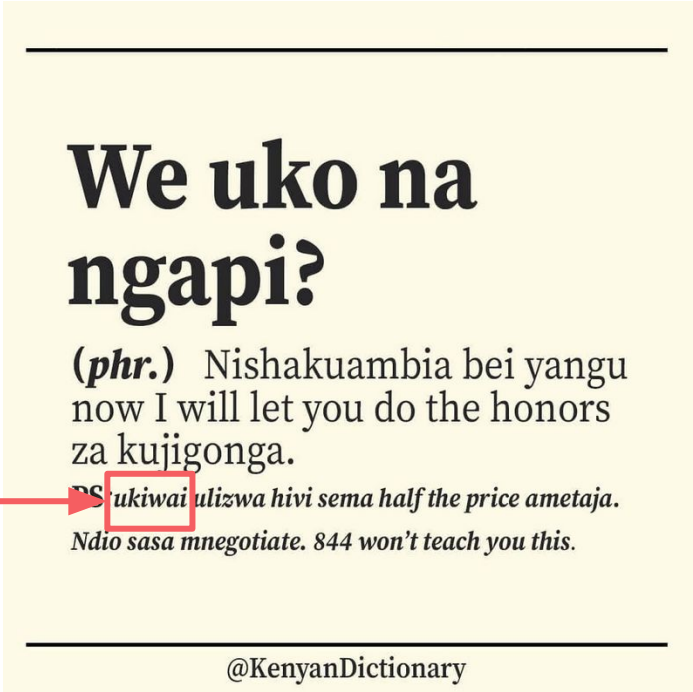
844 (Kenyan education system) won't teach you this.

https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

Challenges faced understanding Kenyans on Twitter

Shorthand
'ukiwai' instead of
'ukiwahi'

English translation:
if you ever



https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

Challenges faced understanding Kenyans on Twitter

Shorthand
'We' instead of
'Wewe'

English translation:
You

https://twitter.com/kenyandictiona/status/1434091356591362051/photo/1

What makes it difficult for a computer to process Kenyan content?

A. Code-switching
B. Non-English words
C. A and B
D. Data format

What makes it difficult for a computer to process Kenyan content?

C. A and B

NLP challenges for understanding Kenyans

- Code-switching
  - Using the same language
  - Using two or more languages
- Slang
  - Combining two words in two different languages to form one word
  - Using an urban dialect
- Multilingualism
  - Mixing words from two or more languages
  - Mixing sentences from two or more languages
- Use of under-resourced languages

We need to create a “living” language model for Kenya.

Project Plan

Have a systematic way of handling code switching for application in other cultures and languages

- Media
- Business
- Social impact

- Manage code-switching
- Preserve context
- Determine if positive or negative

The dataset/corpus

Kenyans’ linguistic identity is hard to define...

… so a representative corpus from Twitter was selected!

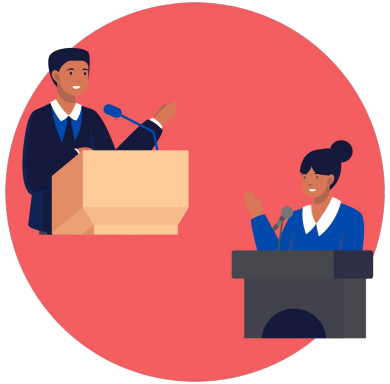
Why Kenyans on Twitter?

Kenyans on Twitter is a decentralized group of Kenyans that use the hashtag #KOT to organize and mobilize around political and socio-cultural issues both within Kenya and outside its borders.

Kenyans may be diverse online, but on Twitter they're self-identifying as one.

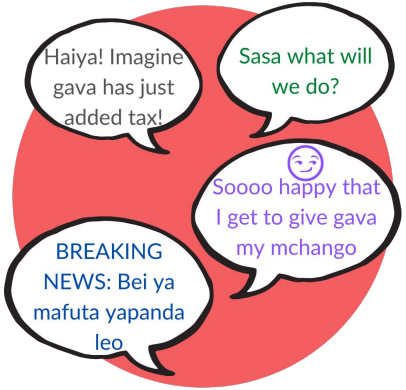
Why a #KOT corpus matters

- A pre-trained corpus can't be used for sentiment analysis of Kenyan speech - there is a high likelihood of code-switching and multilingualism



- Kenyans on Twitter predominantly discuss politics compared to Kenyans on other social media platforms

Why a #KOT corpus matters



- Analyzing large amounts of data with either code-switching or multilingualism with as little translation as possible preserves context

- We need to be able to gauge the sentiment of Kenyan speakers on Twitter

How was the corpus built?

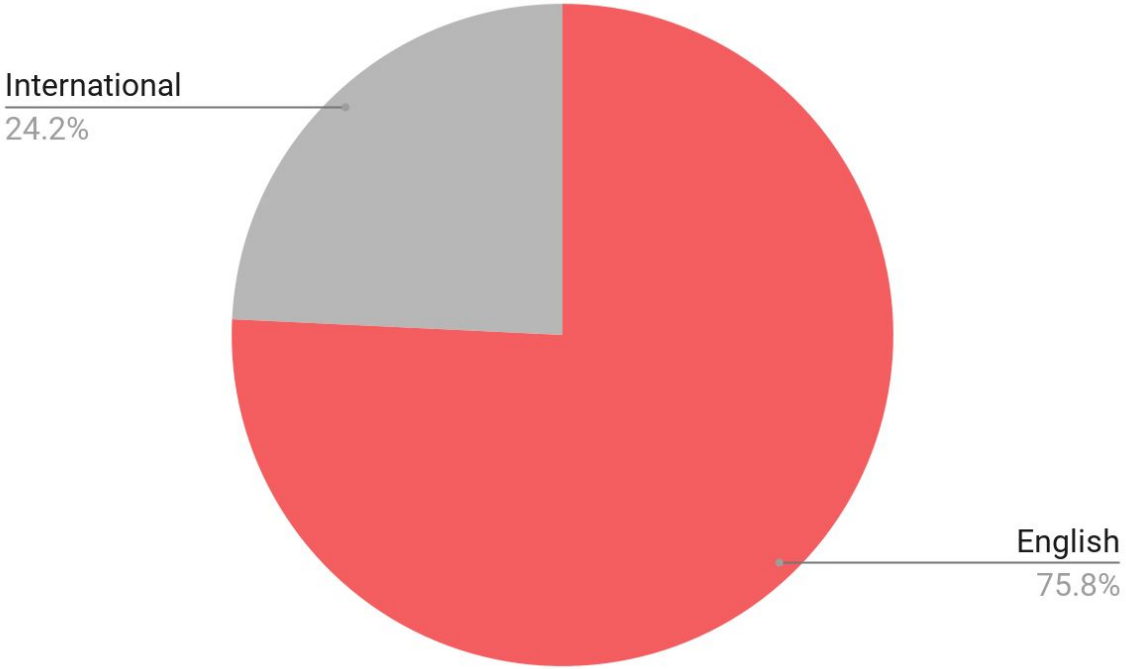
Data Collection

- Tweets collected were for the UTC+3 timezone, with timestamps ranging from 1st June 2020 - 1st June 2021

Data collection was done from a Microsoft command line running

```
twint -s %23kot --since "2020-06-01 03:00:00" --until "2021-06-01 03:00:00" -o kotdata.csv --csv
```

Language distribution in #KOT dataset



Data Pipelining

Initial Access Point

- The data was accessed using a data mining tool via my command line - that way, only relevant data was selected

Data Format

- Versions of the data were then stored into .csv (comma separated value) files, that can be opened in Excel

Data Pipelining

User Interface

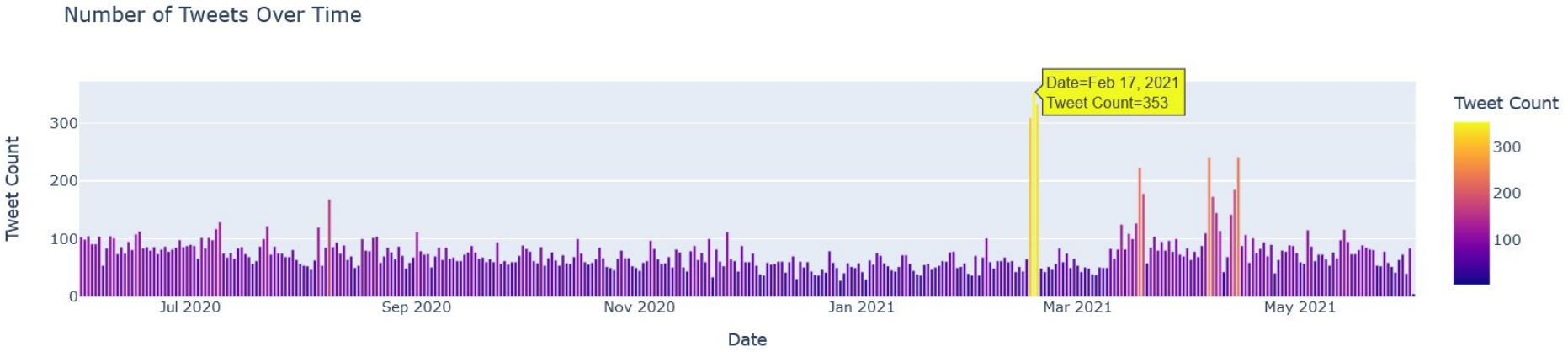
- The data was accessed, modified, visualized and modelled using a Jupyter Notebook
- It ran locally
- It contained blocks of both R and Python code

Tweet Selection

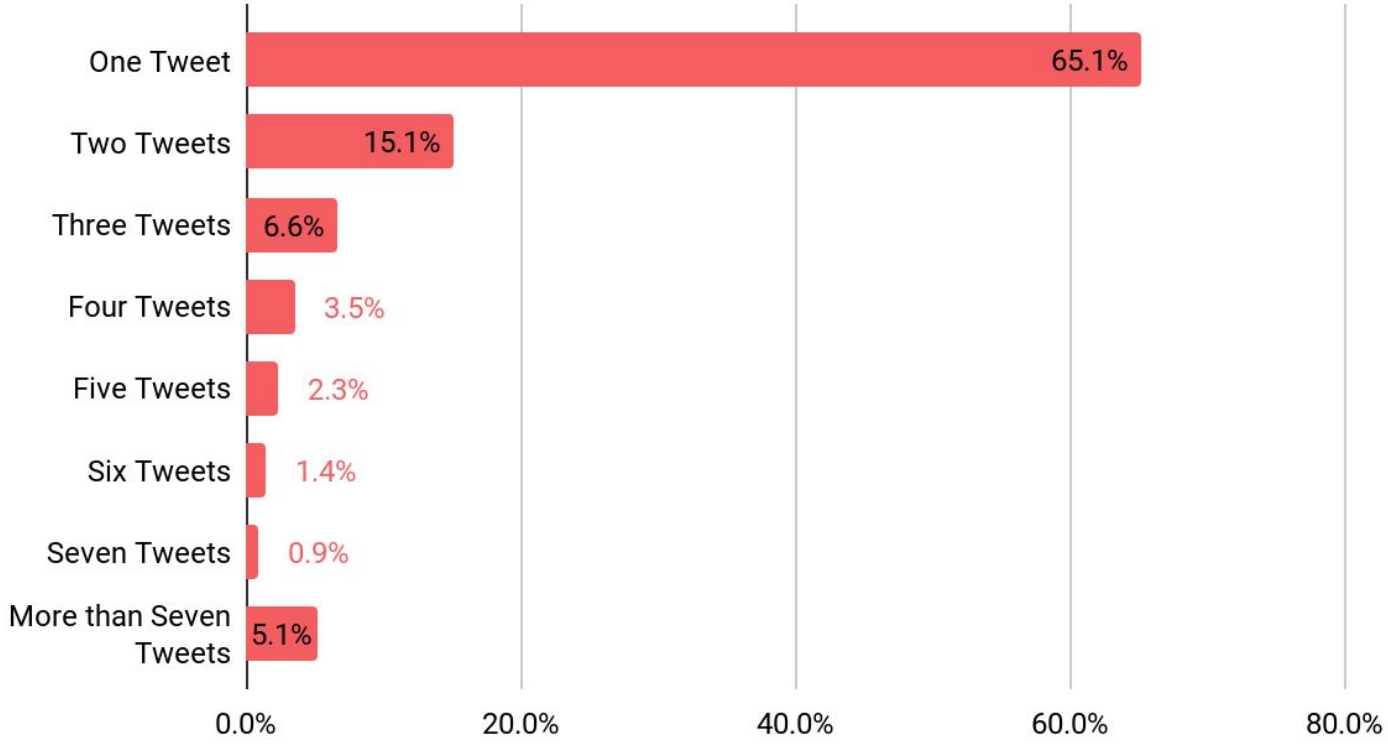
- Tweets selected were tagged ‘en’ by Twitter (English ones)

What characteristics does the corpus have?

Key aspects of the corpus


Number of Tweets Over Time
Date=Feb 17, 2021
Tweet Count=353
Tweet Count
300
200
100
0
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Date
Tweet Count
300
200
100

Number of Tweets per User



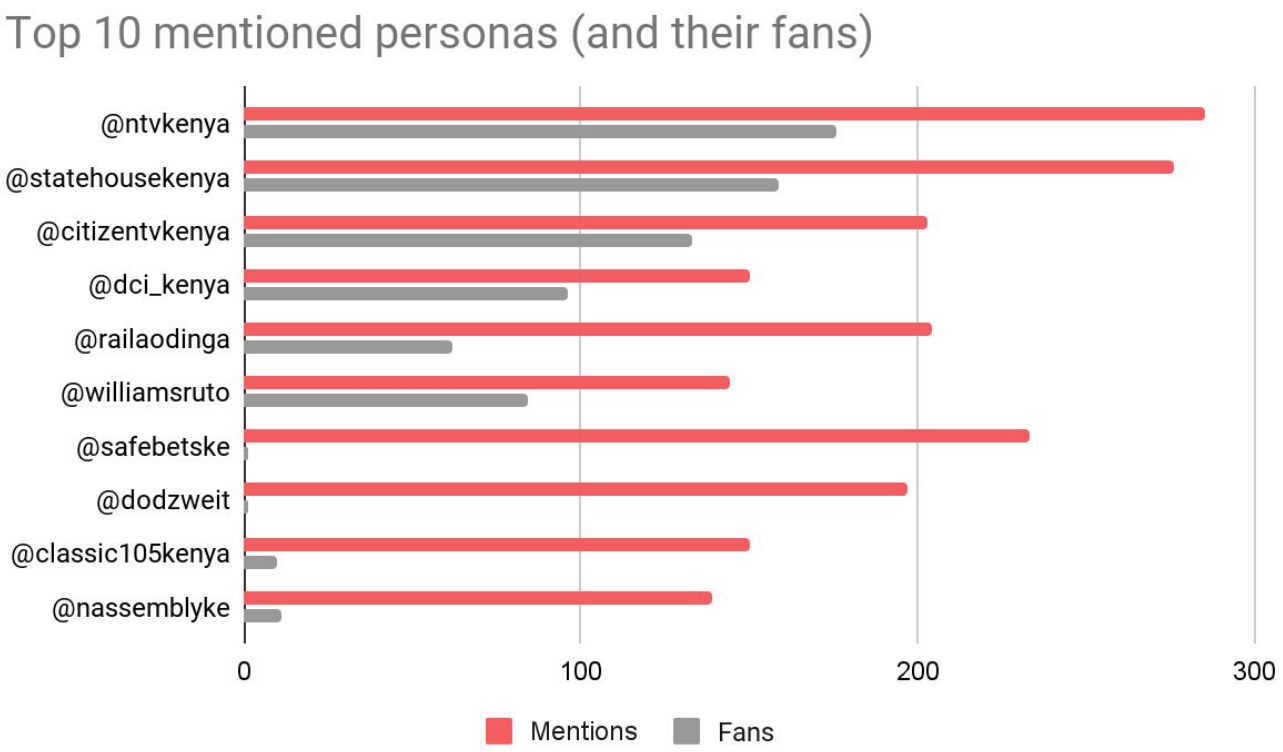
Popular hashtags

Sample "English" tweets

[6] "I don't if you have that friend mwenye huwa mnachat na yeye IG, FB na watsapp at the same time na kwa hzo base zote mnaongea story different #KOT"

[8] "Thank you @BravinYuri for always engaging #KOT on matters #MentalHealthKE I hope one day @MOH_Kenya will see the need to facilitate (finance) these conversations to reach the maaaaaaaany Kenyans who are not online to benefit from the useful information shared here."

[15] "@nsabiyumva_ Who on earth provided translation? God aba #KOT ntimubakira. Yewe ga Yee"


Top 10 mentioned personas (and their fans)
@ntvkenya
@statehousekenya
@citizentvkenya
@dci_kenya
@railaodinga
@williamsruto
@safebetske
@dodzweit
@classic105kenya
@nassemblyke
0
100
200
300
Mentions
Fans

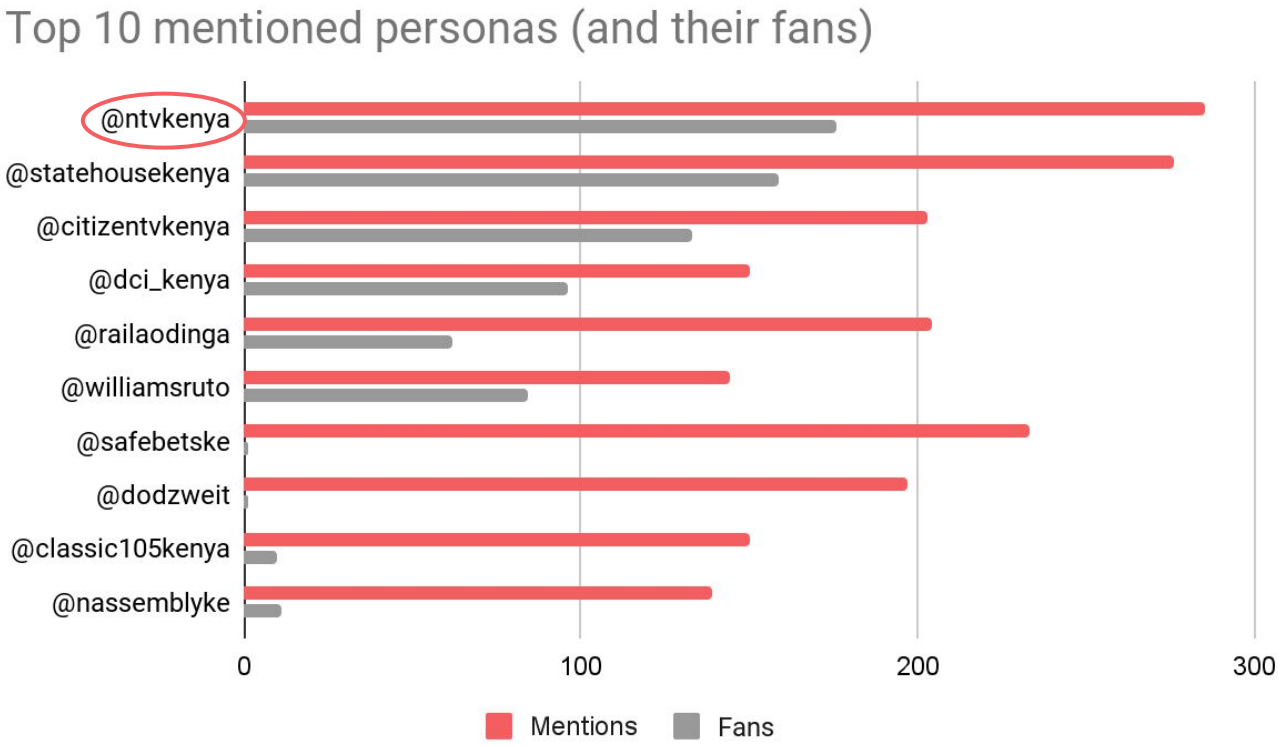
Are Political and Media Personas popular on Kenyan Twitter?

A. No
B. Yes

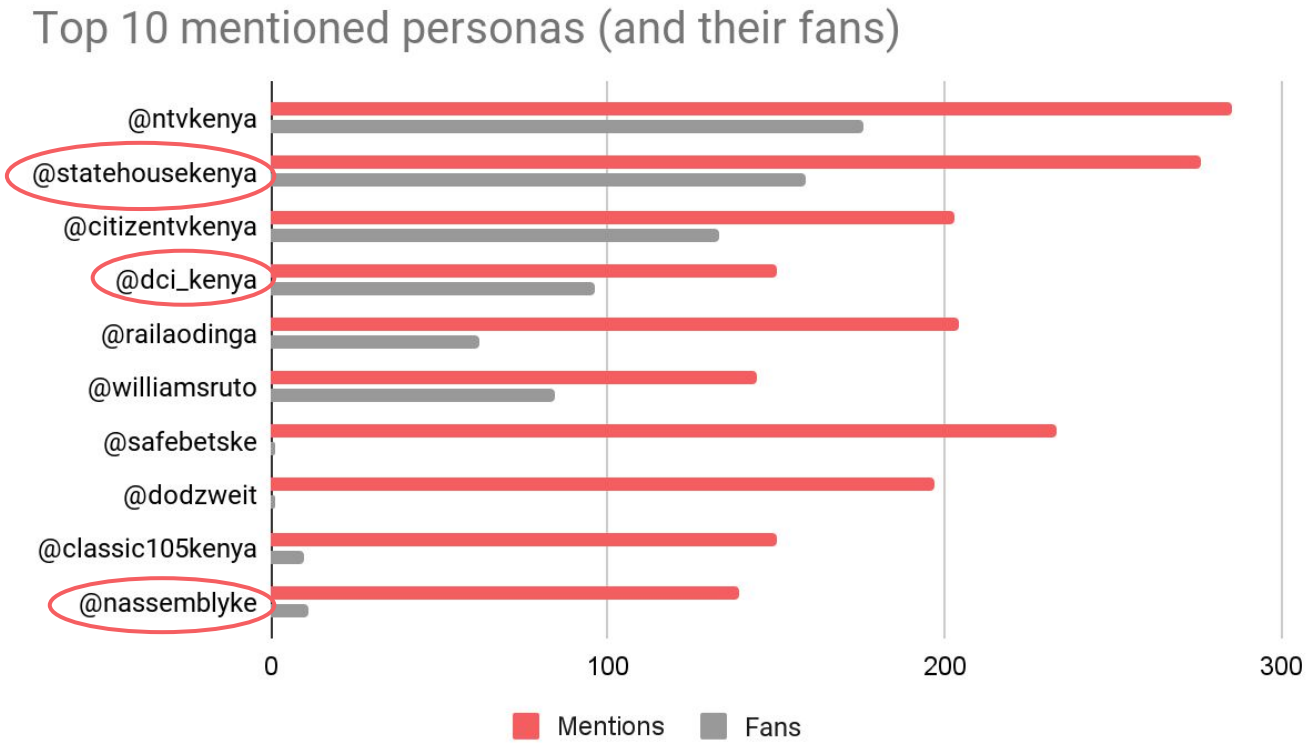
Are Political and Media Personas popular on Kenyan Twitter?

B. Yes

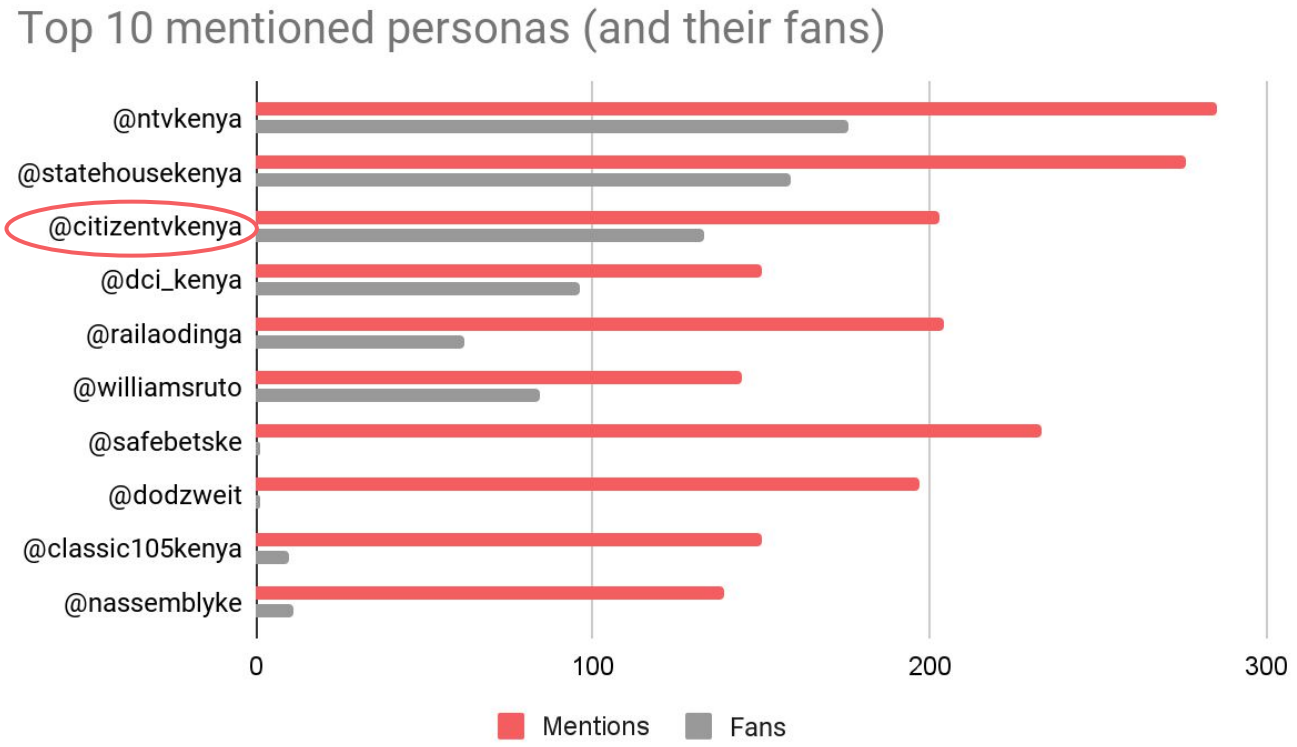
TV station


Top 10 mentioned personas (and their fans)
@ntvkenya
@statehousekenya
@citizentvkenya
@dci_kenya
@railaodinga
@williamsruto
@safebetske
@dodzweit
@classic105kenya
@nassemblyke
0
100
200
300
Mentions
Fans

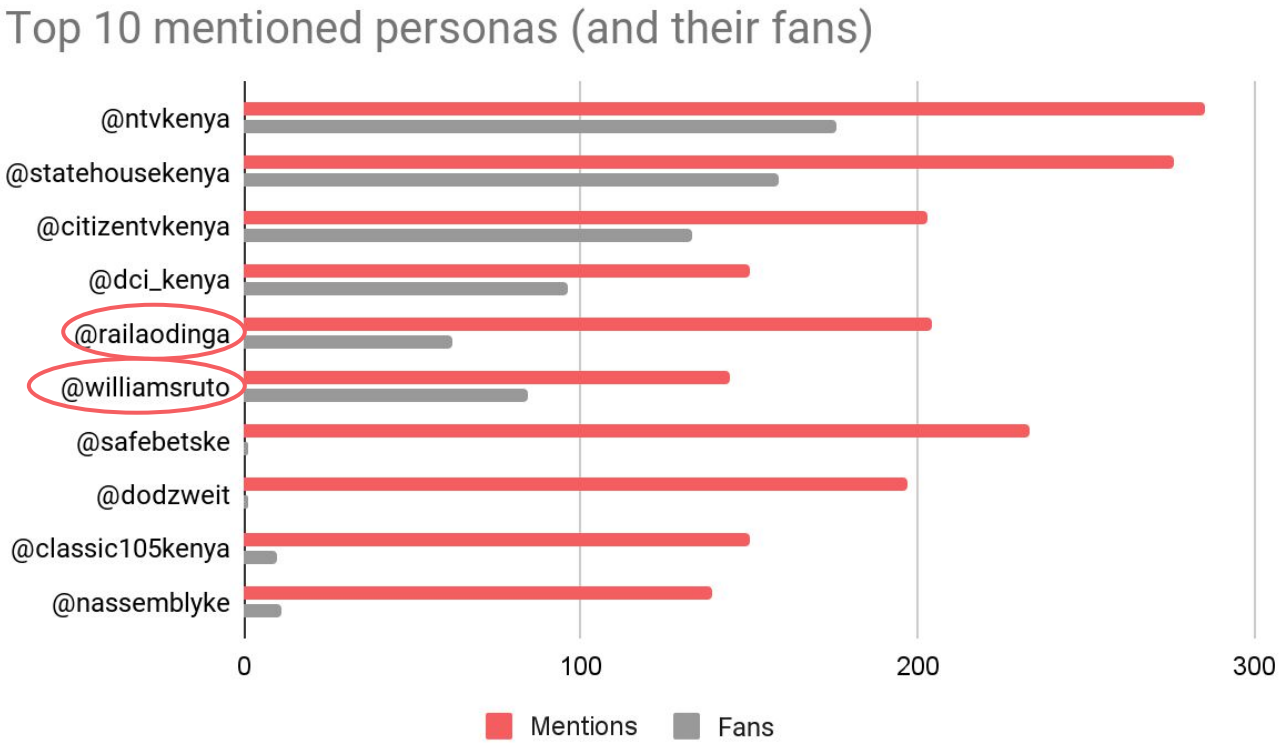
Government agencies



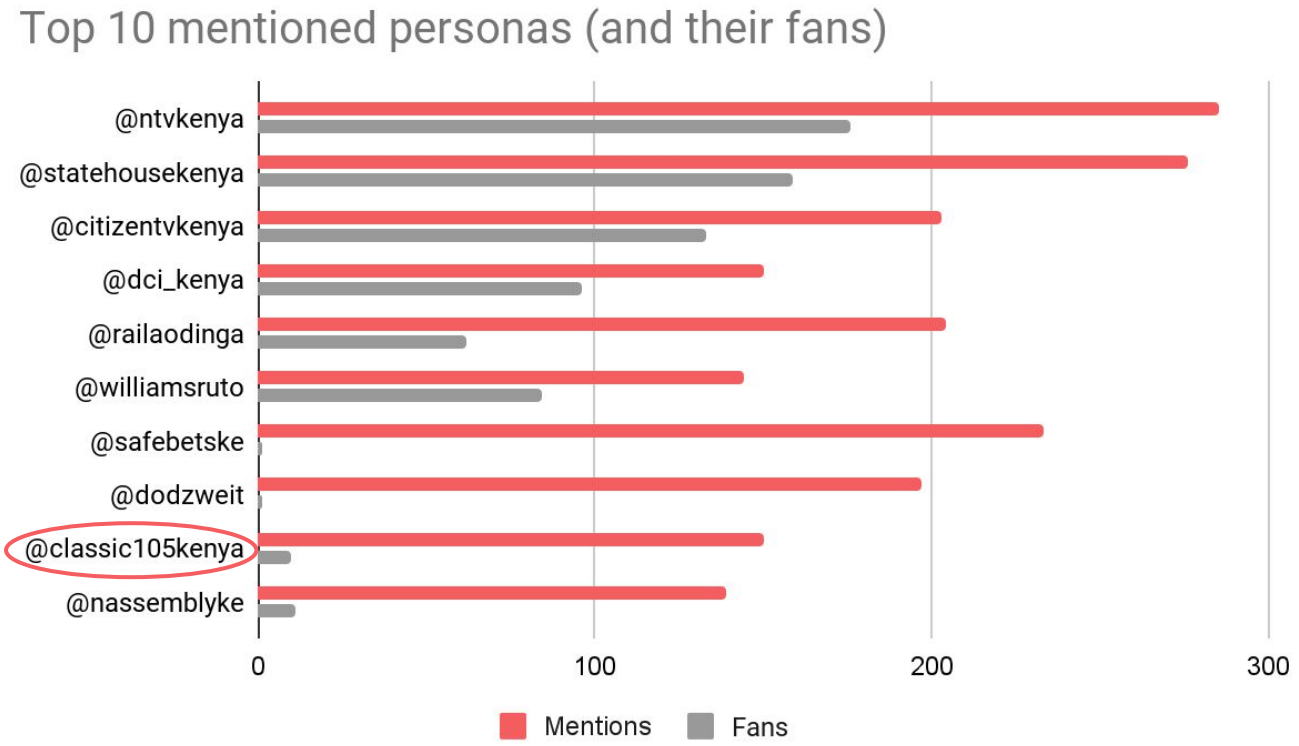
TV station under the same media house controlling Citizen Radio, Top Radio Station in 2019



Seasoned politicians and prospective 2022 Presidential Candidates

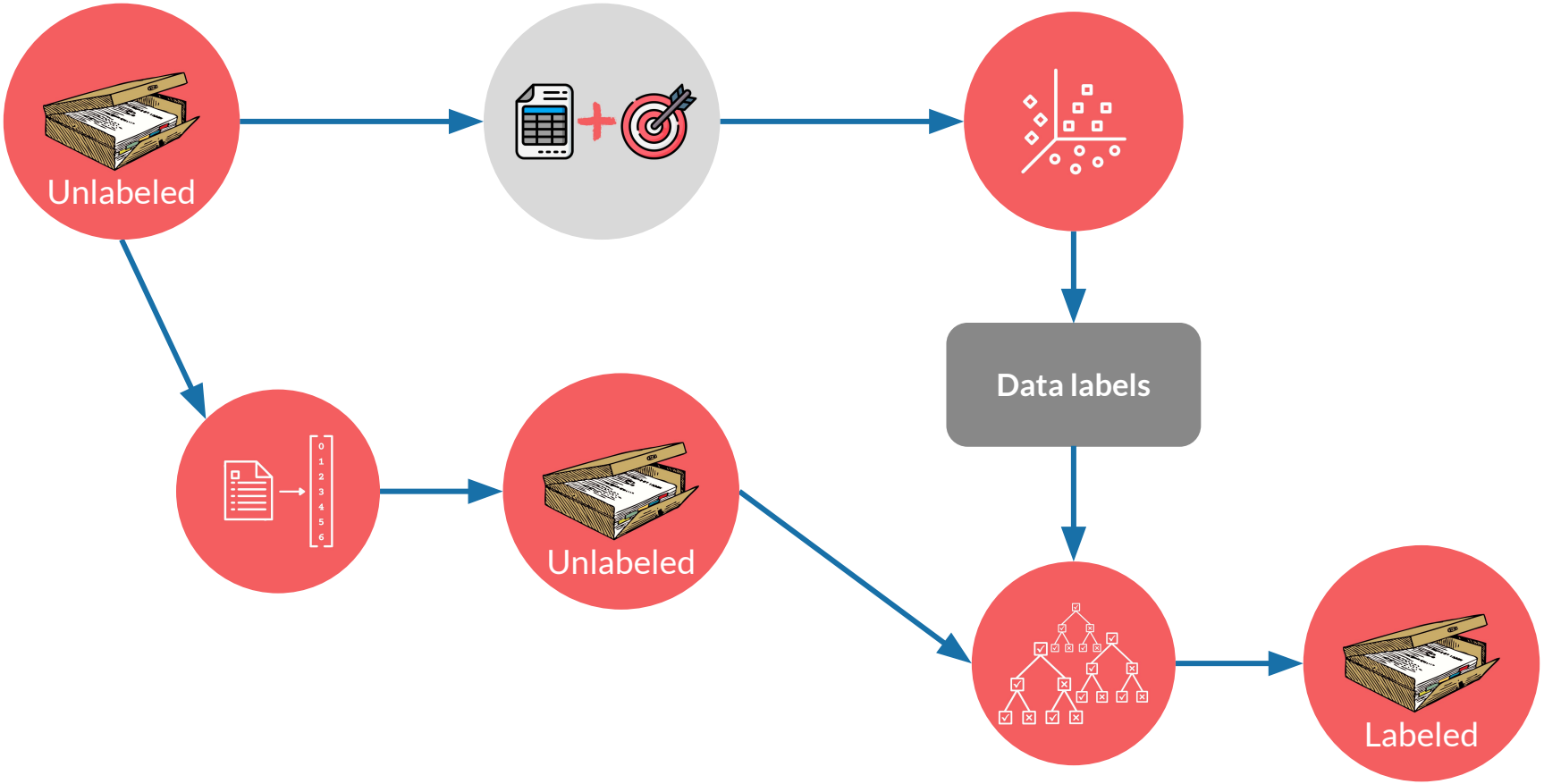


Top 14 English Language Radio Station in 2019



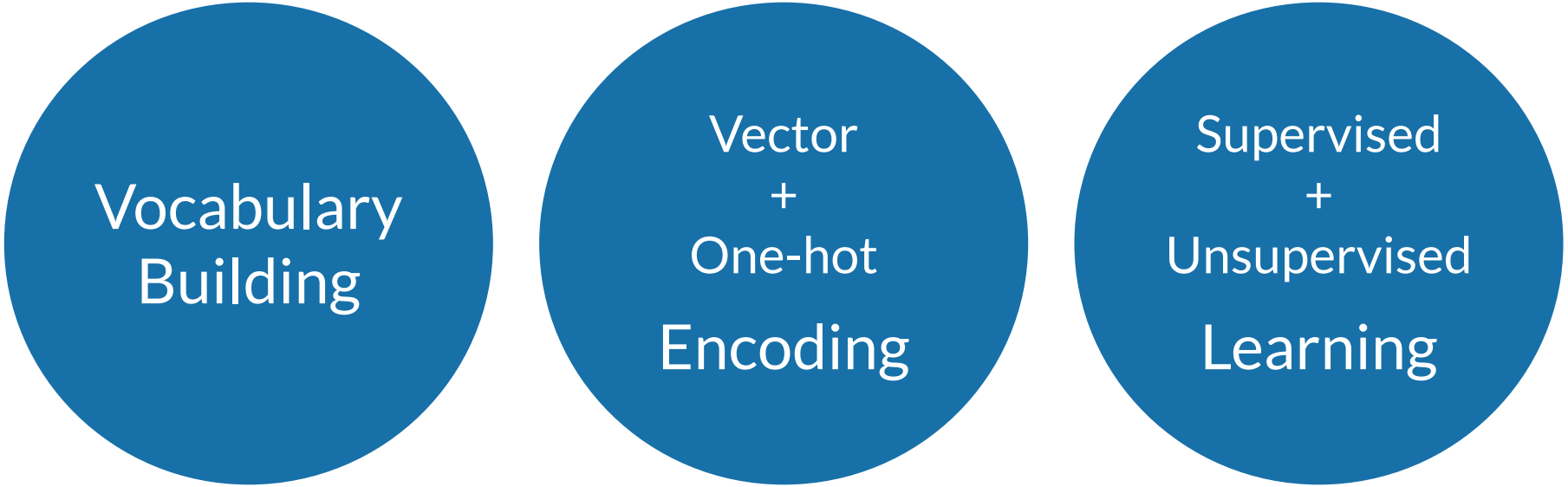
Building the solution

We need to create a “living” language model for Kenya.


Unlabeled
Data labels
Unlabeled
Labeled

How do we create this model?

Key model ingredients



Vocabulary Building

- Infer vectors only from the words in the corpus
- The languages and code-switching become a new “language” to learn

Encoding

- Vector encoding preserves word order and tweet uniqueness
- One-hot encoding identifies tweet sentiment

Machine Learning

- Vectorizes the tweets
- Learns then predicts sentiment
- Labels each tweet as either positive or negative sentiment

Making different forms of the corpus helps the computer understand Kenyans.

A. True
B. False

Making different forms of the corpus helps the computer understand Kenyans.

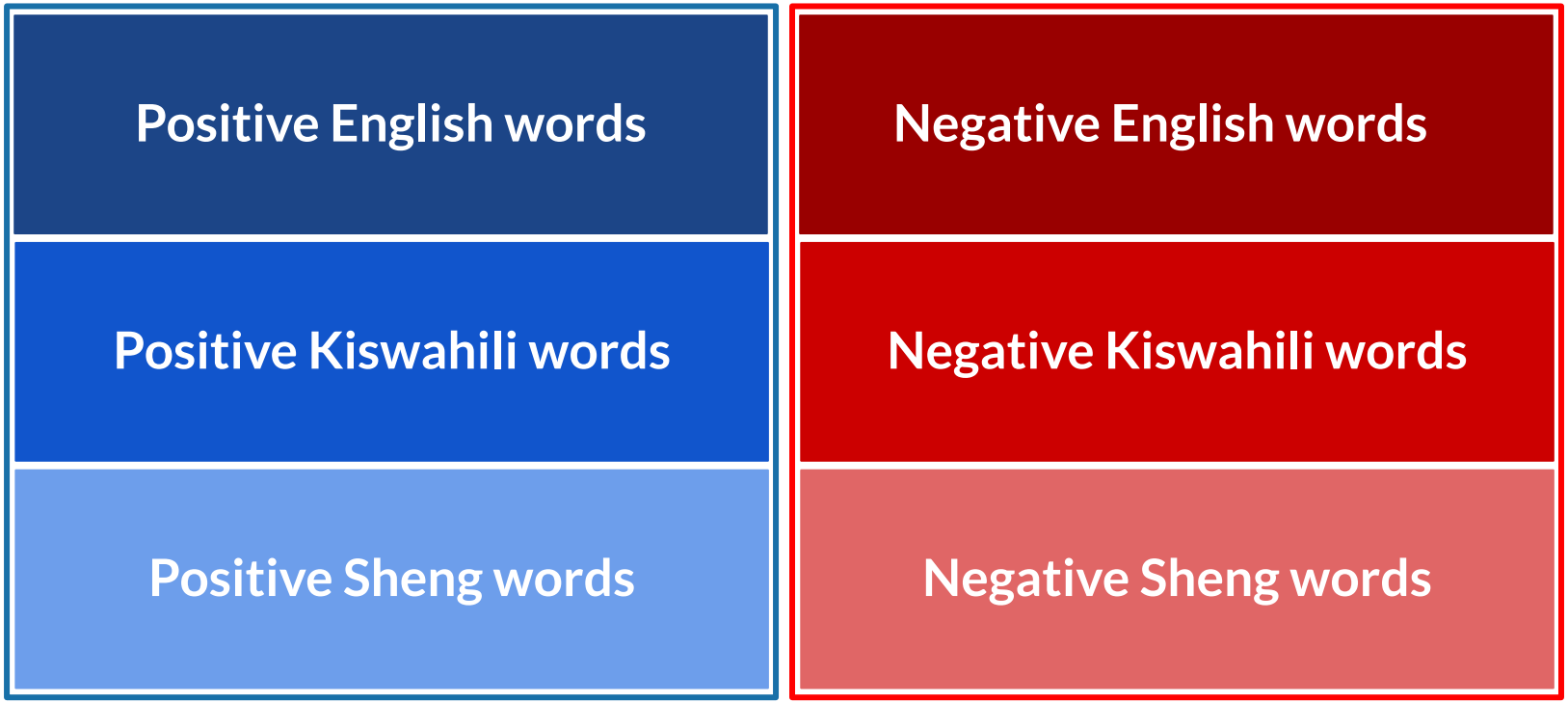
A. True

Vocabulary building

- All tweets
  - Were converted into UTF-8 format
  - Were made lowercase
  - Had punctuation removed, with full stops and commas replaced by spaces
  - Were tokenized and tagged
  - Were vectorized to retain original meaning (each vector had 32 elements)

    only based on the vocabulary in the corpus

One-hot encoding



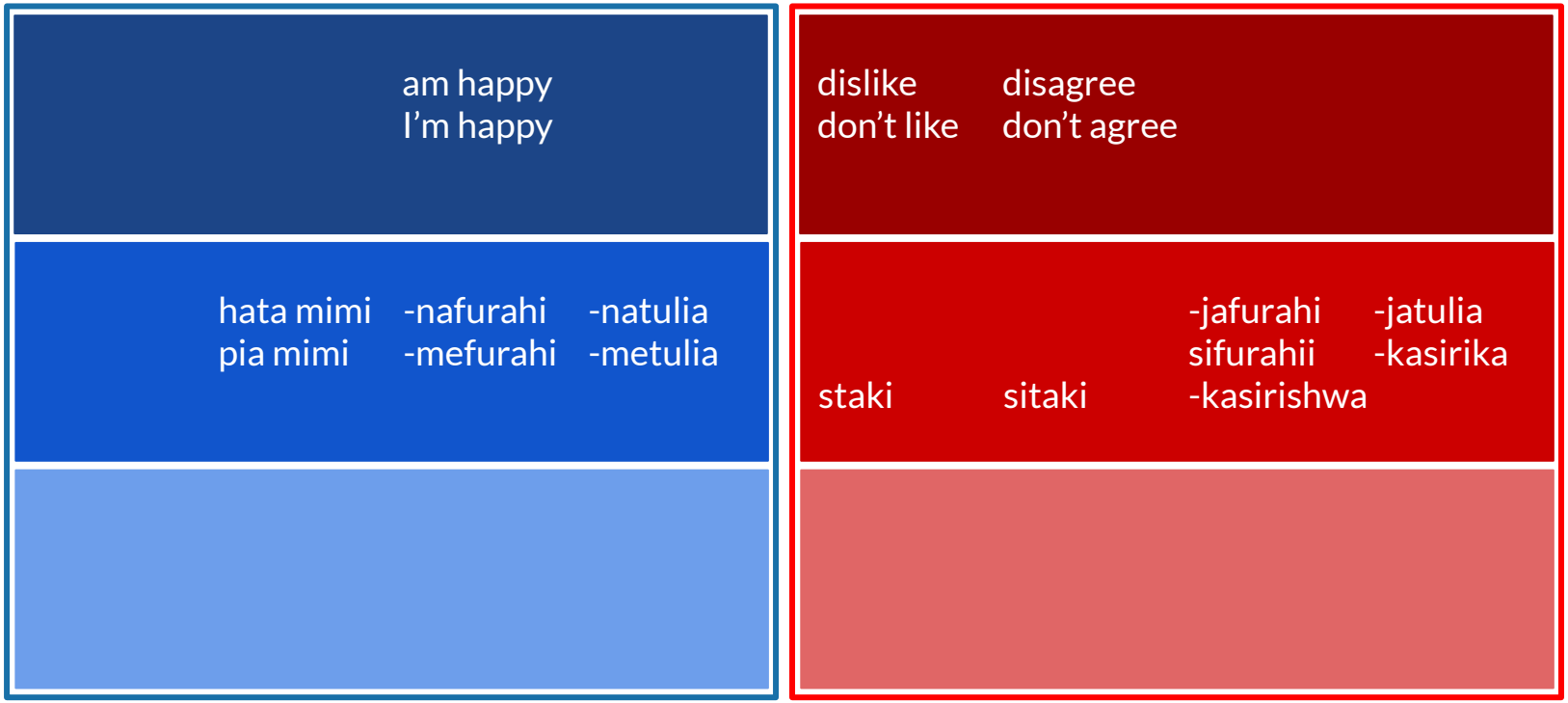
One-hot encoding

like	agree	am happy	grateful
love	even me	I'm happy	okay
I can	want	peace	thankful
-penda	hata mimi	-nafurahi	-natulia
-weza	pia mimi	-mefurahi	-metulia
sawasawa	shukuru	shukran	asante
niko easy			
ni poa			

dislike	disagree	angry	
don't like	don't agree	frustrated	
I can't	don't want	annoyed	
-pendi	hapana	-jafurahi	-jatulia
siwezi	-kataa	sifurahii	-kasirika
staki	sitaki	-kasirishwa	
I kent	si poa		
ngori	nimejam		
siko sawa			

One-hot encoding: Stemmed words

One-hot encoding: Forms of a word

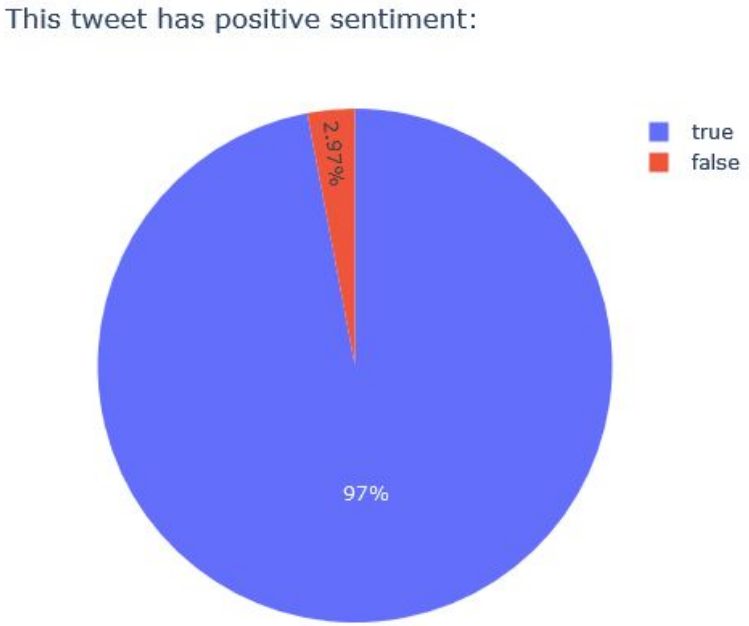


One-hot encoding

```
kmeans_data2 = kmeans_data2.astype('int16')
kmeans_data2['positive'] = sum([kmeans_data2['like'], kmeans_data2['love'], kmeans_data2['penda'], kmeans_data2['agree'],
                                kmeans_data2['even me'], kmeans_data2['hata mimi'], kmeans_data2['pia mimi'],
                                kmeans_data2['i can'], kmeans_data2['weza'], kmeans_data2['niko easy'],
                                kmeans_data2['sawasawa'], kmeans_data2['ni poa'], kmeans_data2['want'],
                                kmeans_data2['am happy'], kmeans_data2['im happy'], kmeans_data2['nafurahi'],
                                kmeans_data2['mefurahi'], kmeans_data2['peace'], kmeans_data2['natulia'],
                                kmeans_data2['metulia'], kmeans_data2['grateful'], kmeans_data2['okay'],
                                kmeans_data2['thankful'], kmeans_data2['shukuru'], kmeans_data2['shukran'],
                                kmeans_data2['asante'] ])
kmeans_data2['is_positive'] = kmeans_data2['positive'] > 0

kmeans_data2['negative'] = sum([kmeans_data2['dislike'], kmeans_data2['dont like'], kmeans_data2['pendi'],
                                kmeans_data2['disagree'], kmeans_data2['dont agree'], kmeans_data2['i kent'],
                                kmeans_data2['i cant'], kmeans_data2['siwezi'], kmeans_data2['ngori'],
                                kmeans_data2['siko sawa'], kmeans_data2['si poa'], kmeans_data2['dont want'],
                                kmeans_data2['staki'], kmeans_data2['sitaki'], kmeans_data2['hapana'],
                                kmeans_data2['kataa'], kmeans_data2['jafurahi'], kmeans_data2['sifurahii'],
                                kmeans_data2['angry'], kmeans_data2['frustrated'], kmeans_data2['annoyed'],
                                kmeans_data2['nimejam'], kmeans_data2['kasirika'], kmeans_data2['kasirishwa'],
                                kmeans_data2['jatulia'] ])
kmeans_data2['is_negative'] = kmeans_data2['negative'] > 0

kmeans_data2['score'] = kmeans_data2['positive'] - kmeans_data2['negative']

print(kmeans_data2.shape)
kmeans_data2.columns
```

Only 4,038 tweets had any sentiment!

Data Validation



We feature engineer the corpus...

Features Shortlisted

- `user_id` : Identifies a specific user of a Twitter account
- `conversation_id` : Identifies a specific conversation in Twitter
- `is_positive` : Identifies **if** the number of positive words > 0
- `is_negative` : Identifies **if** the number of negative words > 0
- `score` : Is the difference between the number of positive and number of negative words
- `tweet_vector` : a 32-length vector encoding each tweet leading to a (27504, 32), two-dimensional array

… and represent the corpus in different ways!

Corpus Representation

	is_positive	is_negative	score	user_id	conversation_id
11	True	False	1	1.322146e+18	1.399421e+18
16	True	False	1	1.164366e+18	1.399383e+18
17	True	False	2	9.832390e+17	1.399382e+18
23	True	False	1	8.618734e+17	1.399345e+18
25	True	True	1	9.493538e+17	1.399343e+18
...	...	...	...	...	...
27482	True	False	1	1.123138e+18	1.267354e+18
27490	True	False	1	2.154684e+09	1.267343e+18
27496	True	False	1	7.188911e+17	1.267323e+18
27502	True	False	1	3.508761e+08	1.267291e+18
27503	True	False	1	1.233154e+18	1.267293e+18

4038 rows × 5 columns

Corpus Representation

	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	-0.054221	-0.059695	0.204735	-0.012434	-0.218431	-0.526059	0.245582	0.352808	-0.084664	0.123656	...	0.268182	0.235359	0.182883	-0.242070	-0.291538	0.198136	0.050008	-0.308937	-0.083188	-0.22083
1	0.214223	0.010163	0.175709	0.158398	0.214602	-0.148901	-0.074405	0.285993	0.277494	0.077204	...	0.014069	0.014350	0.480924	-0.168716	-0.064869	0.013319	-0.210251	-0.123573	-0.081802	-0.22449
2	0.135844	-0.094046	-0.167176	0.335372	-0.312349	-0.171240	0.845063	0.163185	-0.169089	-0.189090	...	0.092265	0.192477	-0.108620	0.131853	-0.505302	-0.293672	0.058623	-0.420413	-0.371532	-0.27368
3	-0.200342	0.038055	0.115383	-0.059901	0.178770	-0.356353	0.019747	0.288630	-0.160592	-0.290369	...	0.337513	-0.036276	0.101864	-0.023040	-0.684704	0.432525	-0.078744	0.312283	-0.339912	0.06842
4	-0.058449	-0.026850	-0.117449	0.045585	-0.392766	-0.476494	0.220576	0.178286	-0.317181	-0.269473	...	0.013432	-0.326399	0.256619	-0.203639	-0.203318	0.089847	-0.237512	-0.160940	-0.015498	-0.25129
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
27499	-0.019238	-0.068312	0.499426	-0.043889	-0.508839	0.098265	0.621087	-0.125135	0.020927	-0.125791	...	0.779029	-0.252367	0.573764	-0.108579	-0.338787	0.010278	0.227661	-0.386283	-0.083465	0.29409
27500	0.414353	-0.306572	0.255197	-0.001606	-0.235279	-0.037229	0.194979	-0.019259	0.078354	-0.028766	...	0.062838	-0.276932	0.467898	-0.077617	-0.282830	-0.105535	0.122426	-0.238071	-0.232538	0.40752
27501	0.240456	-0.027379	0.200210	0.143987	-0.071877	-0.356911	0.765408	0.278083	-0.195484	-0.296828	...	0.249877	0.000901	0.232756	-0.157825	-0.490351	0.455189	0.027652	-0.468572	-0.572950	-0.23071
27502	-0.170253	-0.009599	-0.060626	0.022670	-0.060218	-0.262227	0.018374	-0.070432	-0.004010	0.104702	...	-0.012856	-0.208969	0.203599	-0.005906	0.005790	0.103532	0.033010	-0.175217	-0.143680	-0.01156
27503	-0.019464	-0.003142	0.103369	0.412109	-0.300952	-0.058931	0.178910	-0.114324	0.115652	-0.065321	...	-0.198244	-0.416883	-0.232642	-0.487085	-0.328901	0.689087	0.129943	0.101369	-0.143601	0.23757

27504 rows × 32 columns

Data Relevancy Check

- ☑ Is this data preserving context?
- ☑ Is code-switching managed in this data?
- ☑ Can we determine the sentiment of this data?

Let’s get modelling!

Algorithms

Distributed Memory Model of Paragraph Vectors (PV-DM) Algorithm

This algorithm

- Builds word vectors out of each tweet
- Adds a unique token to each tweet
- Predicts the word in the middle
- Creates a new vector representing the document identity

Optimization function: Gradient Descent

Doc2Vec

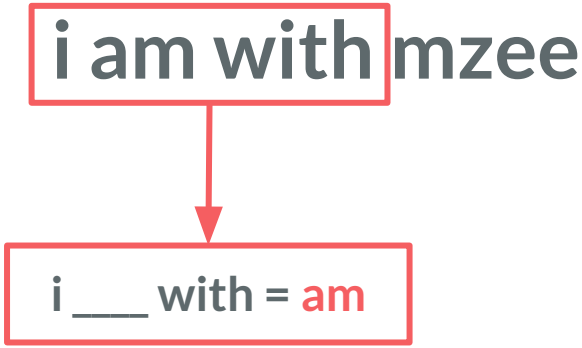
https://arxiv.org/pdf/1405.4053.pdf
https://medium.com/analytics-vidhya/best-nlp-algorithms-to-get-document-similarity-a5559244b23b
https://gab41.lab41.org/doc2vec-to-assess-semantic-similarity-in-source-code-667acb3e62d7

Algorithms: Doc2Vec

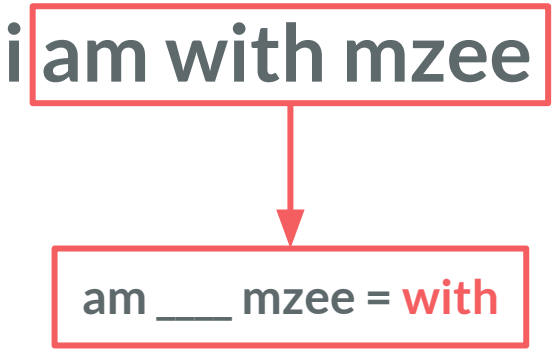
i am with mzee = i + am + with + mzee = `['i', 'am', 'with', 'mzee']`

```
1   0   0   0          1 0 0 0
0   1   0   0          0 1 0 0
0   0   1   0          0 0 1 0
0   0   0   1          0 0 0 1
```

Algorithms: Doc2Vec



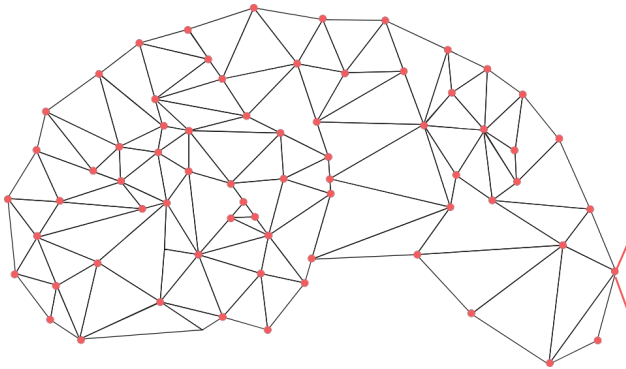
Algorithms: Doc2Vec

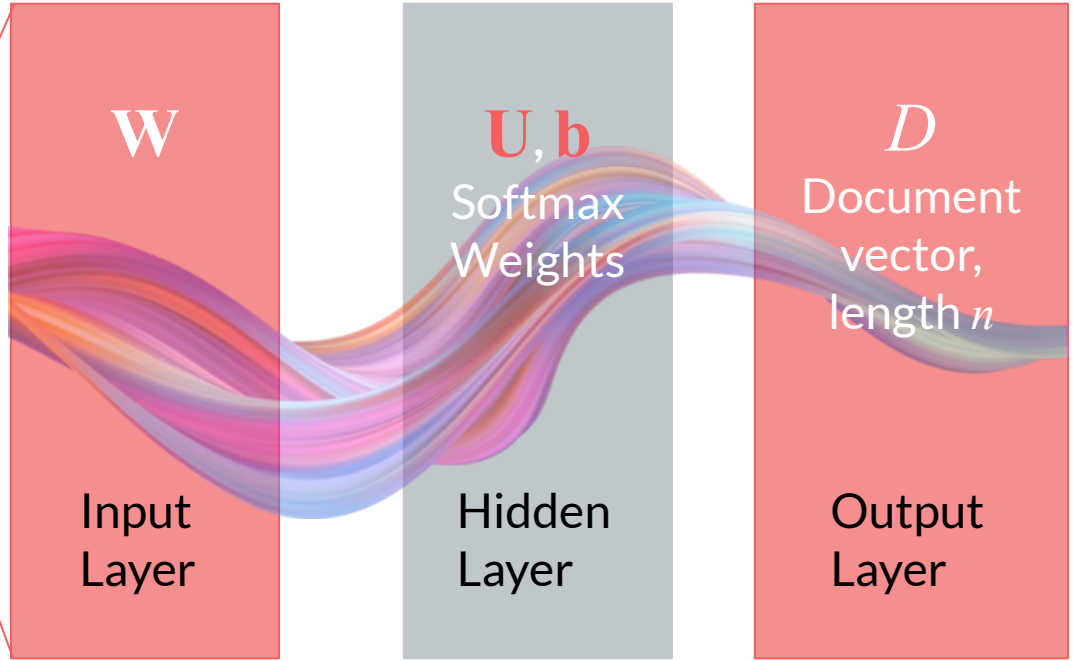


Algorithms: Doc2Vec

$$W = \texttt{['i', 'am', 'with', 'mzee']} = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}$$

Algorithms: Doc2Vec





Vectorized Tweet

```
[1] "@ManUnitedZone_ Made in the streets of mathare,Nairobi,kenya . He was nicknamed Kenyan Messi.Pride of kenya .Friends to @LasetoAbdaz @babazubeyyr
@VictorWanyama #KOT.I insist,The First Kenyan to play for Manchester United .you'll land in trouble for misinforming the masses "
```

```
array([-0.11151993, -0.16104029, -0.03388931,  0.0402423 ,  0.20630357,
       -0.09965932, -0.5022864 , -0.0068965 , -0.16544345, -0.14585833,
        0.18370181, -0.45687547,  0.39285082, -0.10623908, -0.658273  ,
       -0.31019405, -0.11930442,  0.5406505 ,  0.48781177,  0.16552232,
        0.30329403,  0.05520329,  0.4639009 , -0.41770318,  0.40722868,
       -0.00082968, -0.50269216,  0.32031843, -0.49336046,  0.22397785,
       -0.48832273,  0.20908414], dtype=float32)
```

How does the Doc2Vec algorithm work?

A. It encodes the text into word vectors
B. It encodes the text into document vectors
C. It assigns numbers to tweets

How does the Doc2Vec algorithm work?

B. It encodes the text into document vectors

Algorithms

K-means Algorithm

After choosing K number of clusters, this algorithm

- Picks K number of starting points, called centroids
- Finds the Euclidean distance between each point and each centroid

https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/
https://towardsdatascience.com/k-means-clustering-for-beginners-2dc7b2994a4

Algorithms

- Assigns each data point to the nearest centroid
- Gets the mean of each cluster and makes it the new centroid
- Rinse and repeat
  - For n number of iterations **or**
  - Until the centroids stop changing

https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/
https://towardsdatascience.com/k-means-clustering-for-beginners-2dc7b2994a4

K-means algorithms sorts data using the mean of each group.

A. True
B. False

K-means algorithms sorts data using the mean of each group.

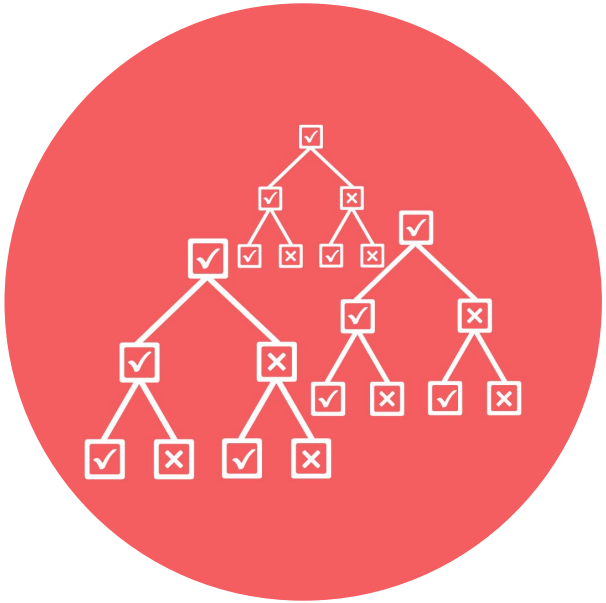
A. True

Algorithms

Ensemble CART Algorithm

This algorithm works as follows:

- A number of CART decision trees answer a series of 'Yes, No' questions with a score along the node of each tree
- The criteria for a binary split can be the error rate or the Gini index
- When the tree hits the maximum depth, a total score is taken of the entire tree



XG Boost

https://arxiv.org/pdf/1603.02754.pdf
https://towardsdatascience.com/xgboost-mathematics-explained-58262530904a
https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

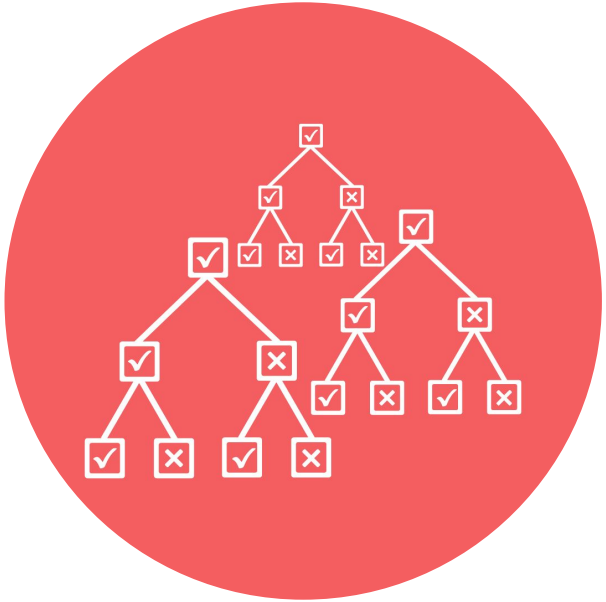
Algorithms

- The ultimate score for the sample is then the sum of all the total scores of each tree
- Get the errors made and then
  - Run another round of the model showing errors of the previous round
  - Rinse and repeat for m number of rounds

Optimization functions:

Regularization (fight overfitting),

Gradient Boosting (improve quality)



XG Boost

https://arxiv.org/pdf/1603.02754.pdf
https://towardsdatascience.com/xgboost-mathematics-explained-58262530904a
https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

How many splits does a decision tree in XG Boost have?

A. No splits
B. 3 splits
C. 5 splits
D. 2 splits

How many splits does a decision tree in XG Boost have?

D. 2 splits

Each algorithm is used differently.

Vocabulary Building

Vector + One-hot Encoding

Supervised + Unsupervised Learning

UNSUPERVISED

- Vocabulary trainer
- Ran each time the corpus has changes
- Vectorizer

UNSUPERVISED

- Data labeller

SUPERVISED

- Data labeller

27504 Raw Tweets

	is_positive	is_negative	score	user_id	conversation_id
11	True	False	1	1.322146e+18	1.399421e+18
16	True	False	1	1.164366e+18	1.399383e+18
17	True	False	2	9.832390e+17	1.399382e+18
23	True	False	1	8.618734e+17	1.399345e+18
25	True	True	1	9.493538e+17	1.399343e+18
...	...	...	...	...	...
27482	True	False	1	1.123138e+18	1.267354e+18
27490	True	False	1	2.154684e+09	1.267343e+18
27496	True	False	1	7.188911e+17	1.267323e+18
27502	True	False	1	3.508761e+08	1.267291e+18
27503	True	False	1	1.233154e+18	1.267293e+18

4038 rows × 5 columns

	0	1	2	3	4	5	6	
0	-0.054221	-0.059695	0.204735	-0.012434	-0.218431	-0.526059	0.245582	0.352
1	0.214223	0.010163	0.175709	0.158398	0.214602	-0.148901	-0.074405	0.285
2	0.135844	-0.094046	-0.167176	0.335372	-0.312349	-0.171240	0.845063	0.163
3	-0.200342	0.038055	0.115383	-0.059901	0.178770	-0.356353	0.019747	0.288
4	-0.058449	-0.026850	-0.117449	0.045585	-0.392766	-0.476494	0.220576	0.178
...	...	...	...	...	...	...	...	
27499	-0.019238	-0.068312	0.499426	-0.043889	-0.508839	0.098265	0.621087	-0.125
27500	0.414353	-0.306572	0.255197	-0.001606	-0.235279	-0.037229	0.194979	-0.019
27501	0.240456	-0.027379	0.200210	0.143987	-0.071877	-0.356911	0.765408	0.278
27502	-0.170253	-0.009599	-0.060626	0.022670	-0.060218	-0.262227	0.018374	-0.070
27503	-0.019464	-0.003142	0.103369	0.412109	-0.300952	-0.058931	0.178910	-0.114

27504 rows × 32 columns

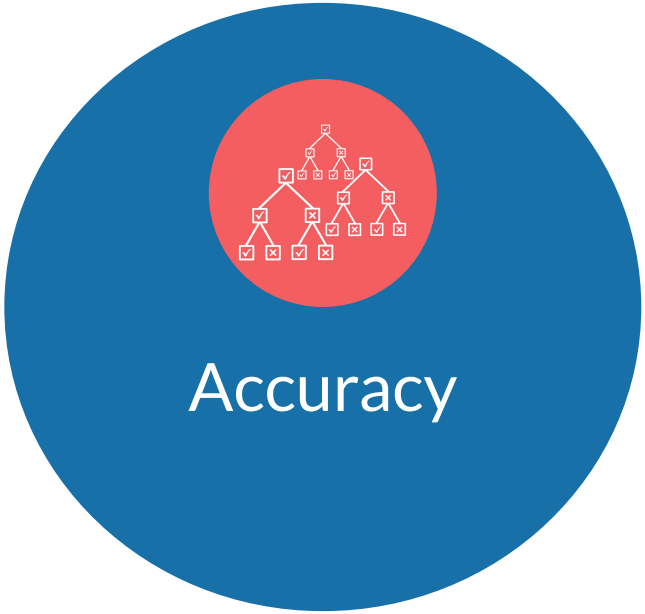
Data labels for 4038 Tweets

27504 Labelled Tweets

Evaluating the solution

How can we know that we have the “best” solution?

Modelling Metrics



0: smackasoreass tobibwy e teezey1286 yes how could you   kot

1: im sure this was a demo video kot kenya africa ufisadi

2: the generation that buys phones that cost 3 times their parents monthly salary are surprised that their parents think that they are working somewhere the parents are just telling them to keep working kot

3: manunitedzone made in the streets of mathare nairobi kenya   he was nicknamed kenyan messi pride of kenya friends to lasetoabdaz babazubeyyr victorwanyama kot i insist the first kenyan to play for manchester united youll land in trouble for misinforming the masses

4: madam gracekuriake we thank you for all the job updates that you give kot you are impacting lives   one job at a time

5: i dont if you have that friend mwenye huwa mnachat na yeye ig fb na watsapp a t the same time na kwa hzo base zote mnaongea story different kot

6: 7 more followers to go kot

7: thank you bravinyuri for always engaging kot on matters mentalhealthke i hop e one day moh kenya will see the need to facilitate finance these conversations to reach the maaaaaaaany kenyans who are not online to benefit from the useful in formation shared here

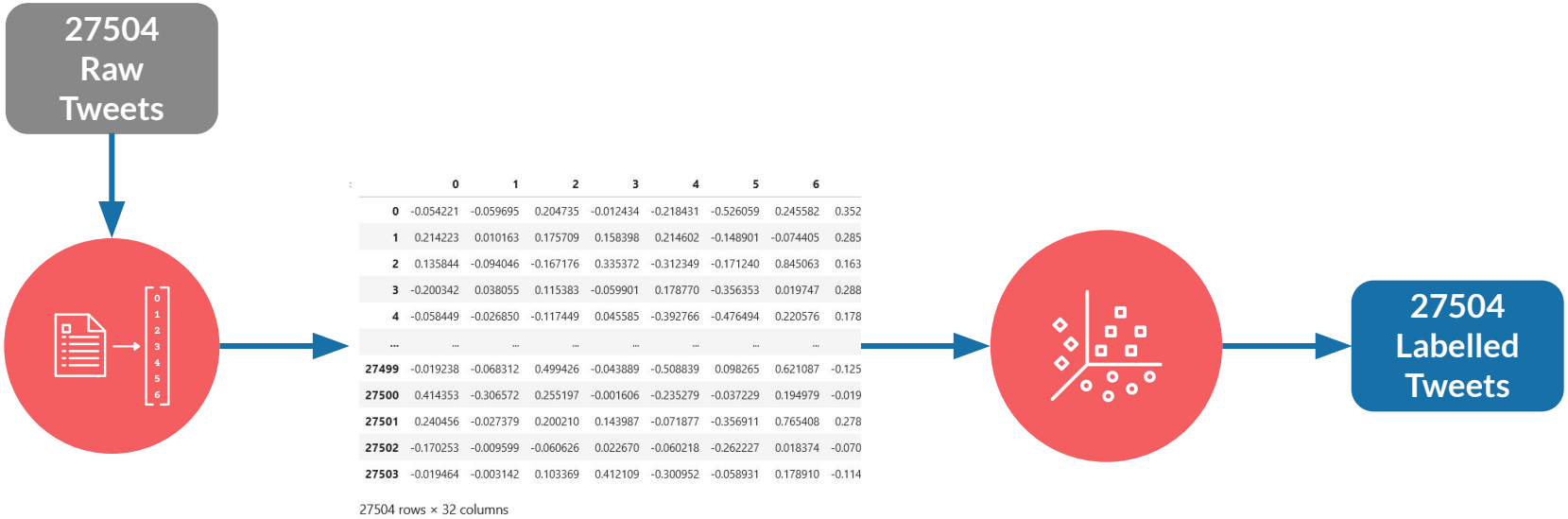
8: check out taking online learning to a whole new level kot ukinkenya usaid education gpforeducation raiseyourhand diani carolinemutoko

9: when you spend thousands of kshs to see queen of the jungle then you spend thousands of kshs to see your queen on royale spring mattress kot ikokazike kead rive kenyans parklands kotsokoni kot nairobi madarakaday2021 airbnb kenyan mattre ss madeinkenya madeinkenya

Human Review

Tweet Index	Sentiment
0	Neutral
1	- (sarcasm)
2	- (sarcasm)
3	+ (humour)
4	+
5	Neutral
6	+
7	+
8	+ (advert)
9	+ (advert)

Model One



	0	1	2	3	4	5	6	
0	-0.054221	-0.059695	0.204735	-0.012434	-0.218431	-0.526059	0.245582	0.352
1	0.214223	0.010163	0.175709	0.158398	0.214602	-0.148901	-0.074405	0.285
2	0.135844	-0.094046	-0.167176	0.335372	-0.312349	-0.171240	0.845063	0.163
3	-0.200342	0.038055	0.115383	-0.059901	0.178770	-0.356353	0.019747	0.288
4	-0.058449	-0.026850	-0.117449	0.045585	-0.392766	-0.476494	0.220576	0.178
...	...	...	...	...	...	...	...	
27499	-0.019238	-0.068312	0.499426	-0.043889	-0.508839	0.098265	0.621087	-0.125
27500	0.414353	-0.306572	0.255197	-0.001606	-0.235279	-0.037229	0.194979	-0.019
27501	0.240456	-0.027379	0.200210	0.143987	-0.071877	-0.356911	0.765408	0.278
27502	-0.170253	-0.009599	-0.060626	0.022670	-0.060218	-0.262227	0.018374	-0.070
27503	-0.019464	-0.003142	0.103369	0.412109	-0.300952	-0.058931	0.178910	-0.114

27504 rows × 32 columns

Human Review

KEY:

- Group 0
- Group 1

Tweet Index	Model One
0	Neutral
1	- (sarcasm)
2	- (sarcasm)
3	+ (humour)
4	+
5	Neutral
6	+
7	+
8	+ (advert)
9	+ (advert)

Model Two

27504 Raw Tweets

	user_idLIST	conversation_idLIST	tweet_vectorLIST
0	1.184135e+18	1.399232e+18	-0.126462
1	1.184135e+18	1.399232e+18	-0.095535
2	1.184135e+18	1.399232e+18	0.203995
3	1.184135e+18	1.399232e+18	-0.027006
4	1.184135e+18	1.399232e+18	-0.315601
...	...	...	...
880123	1.233154e+18	1.267293e+18	0.595895
880124	1.233154e+18	1.267293e+18	-0.148532
880125	1.233154e+18	1.267293e+18	0.065002
880126	1.233154e+18	1.267293e+18	-0.150835
880127	1.233154e+18	1.267293e+18	0.359433

880128 rows × 3 columns

27504 Labelled Tweets

Human Review

KEY:

Group 0

Group 1

Tweet Index	Model One	Model Two
0	Neutral	Neutral
1	- (sarcasm)	- (sarcasm)
2	- (sarcasm)	- (sarcasm)
3	+ (humour)	+ (humour)
4	+	+
5	Neutral	Neutral
6	+	+
7	+	+
8	+ (advert)	+ (advert)
9	+ (advert)	+ (advert)

27504 Raw Tweets

	is_positive	is_negative	score	user_id	conversation_id
11	True	False	1	1.322146e+18	1.399421e+18
16	True	False	1	1.164366e+18	1.399383e+18
17	True	False	2	9.832390e+17	1.399382e+18
23	True	False	1	8.618734e+17	1.399345e+18
25	True	True	1	9.493538e+17	1.399343e+18
...	...	...	...	...	...
27482	True	False	1	1.123138e+18	1.267354e+18
27490	True	False	1	2.154684e+09	1.267343e+18
27496	True	False	1	7.188911e+17	1.267323e+18
27502	True	False	1	3.508761e+08	1.267291e+18
27503	True	False	1	1.233154e+18	1.267293e+18

4038 rows × 5 columns

	0	1	2	3	4	5	6	
0	-0.054221	-0.059695	0.204735	-0.012434	-0.218431	-0.526059	0.245582	0.352
1	0.214223	0.010163	0.175709	0.158398	0.214602	-0.148901	-0.074405	0.285
2	0.135844	-0.094046	-0.167176	0.335372	-0.312349	-0.171240	0.845063	0.163
3	-0.200342	0.038055	0.115383	-0.059901	0.178770	-0.356353	0.019747	0.288
4	-0.058449	-0.026850	-0.117449	0.045585	-0.392766	-0.476494	0.220576	0.178
...	...	...	...	...	...	...	...	
27499	-0.019238	-0.068312	0.499426	-0.043889	-0.508839	0.098265	0.621087	-0.125
27500	0.414353	-0.306572	0.255197	-0.001606	-0.235279	-0.037229	0.194979	-0.019
27501	0.240456	-0.027379	0.200210	0.143987	-0.071877	-0.356911	0.765408	0.278
27502	-0.170253	-0.009599	-0.060626	0.022670	-0.060218	-0.262227	0.018374	-0.070
27503	-0.019464	-0.003142	0.103369	0.412109	-0.300952	-0.058931	0.178910	-0.114

27504 rows × 32 columns

Data labels for 4038 Tweets

27504 Labelled Tweets

Human Review

KEY:

Group 0

Group 1

Tweet Index	Model One	Model Two	Model Three
0	Neutral	Neutral	Neutral
1	- (sarcasm)	- (sarcasm)	- (sarcasm)
2	- (sarcasm)	- (sarcasm)	- (sarcasm)
3	+ (humour)	+ (humour)	+ (humour)
4	+	+	+
5	Neutral	Neutral	Neutral
6	+	+	+
7	+	+	+
8	+ (advert)	+ (advert)	+ (advert)
9	+ (advert)	+ (advert)	+ (advert)

Accuracy:
0.6270

K-Fold Number	Accuracy
1	0.6205
2	0.6246
3	0.6337
4	0.6361
5	0.637
6	0.6246
7	0.6188
8	0.6196
9	0.6246
10	0.6304

What is the “best” solution?

Human Review

Group	Model One	Model Two	Model Three
0	Majorly Neutral	Majorly Positive	Majorly Positive
1	Majorly Positive	Majorly Positive	Majorly Negative

Why was Model Three the best model?

A. It mostly separated positive and negative tweets
B. It was able to handle a very imbalanced dataset
C. It incorporated supervised machine learning
D. All the above

Why was Model Three the best model?

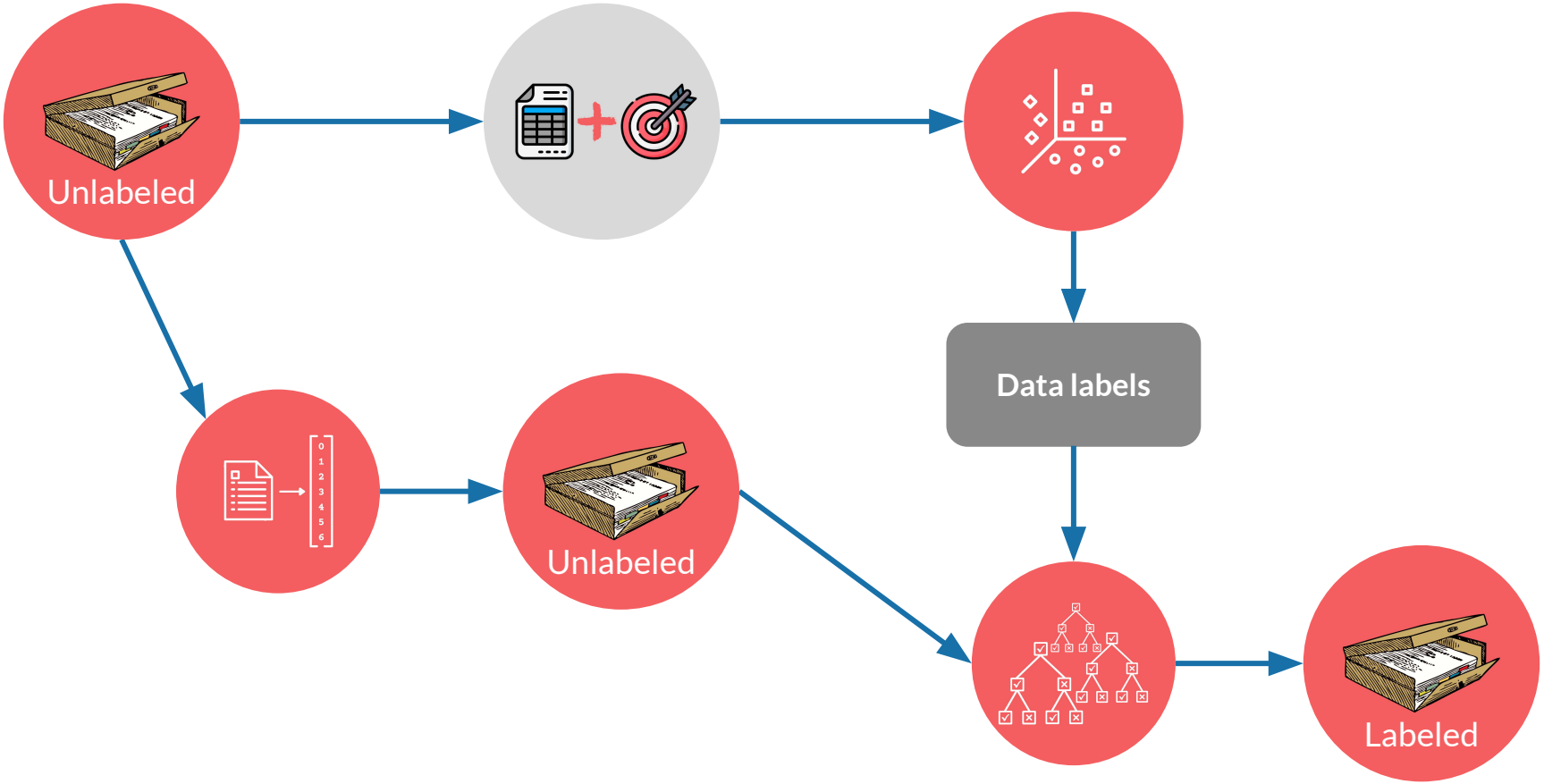
D. All the above

Modelling Insights

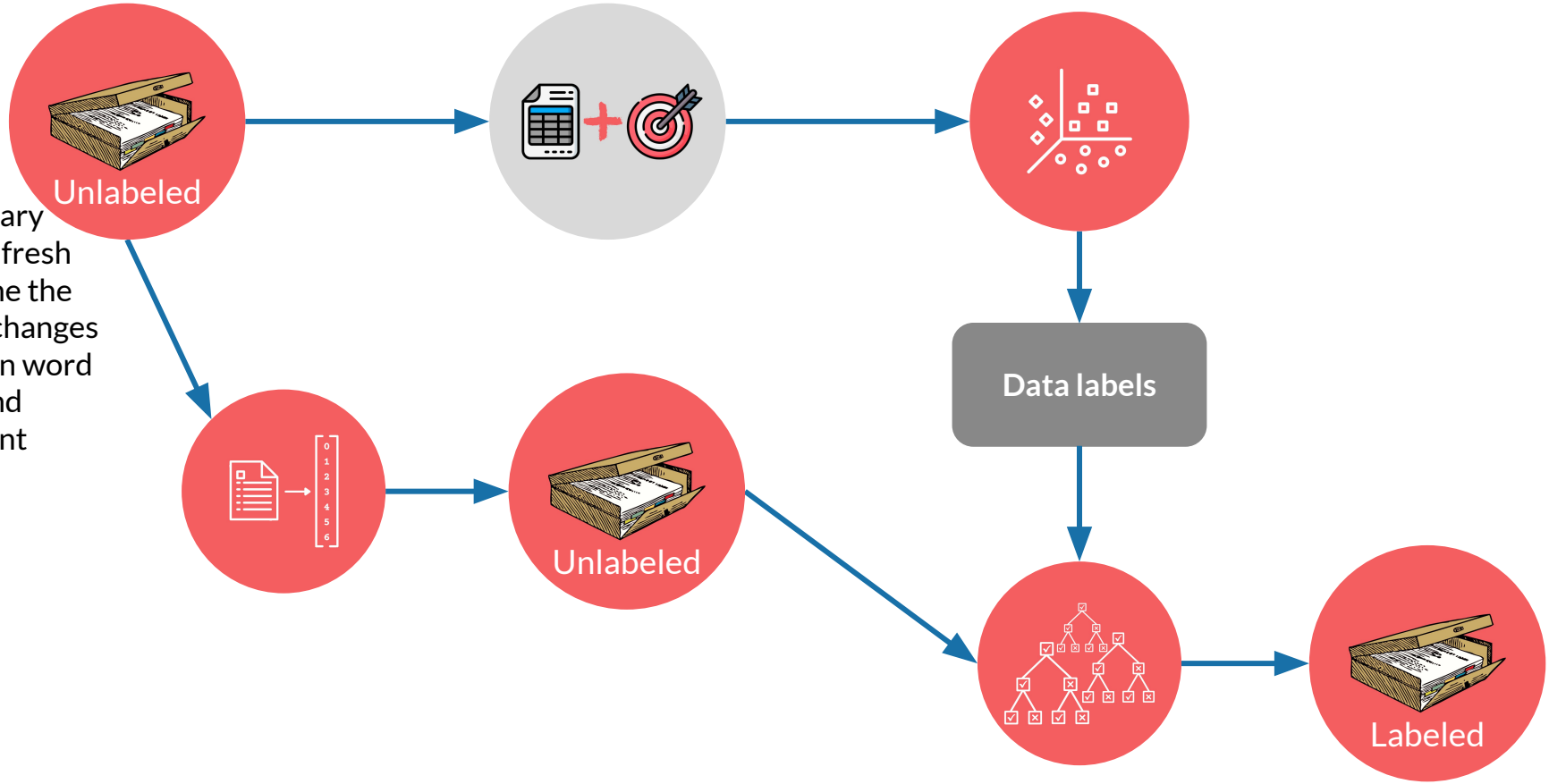
- The classification is not perfect
- It did well for a predominantly unsupervised challenge
- Adverts used popular hashtags not as intended
- Few Swahili and Sheng encoded words were found in the dataset
- The sentiment of most tweets was neutral
- The overall accuracy for the best labelling approach was 62.7%
- Each time the models are re-run, the groupings for 2/3 approaches change except the sample classifications in the third approach, even after setting a random seed

Benefits of the #KOT model

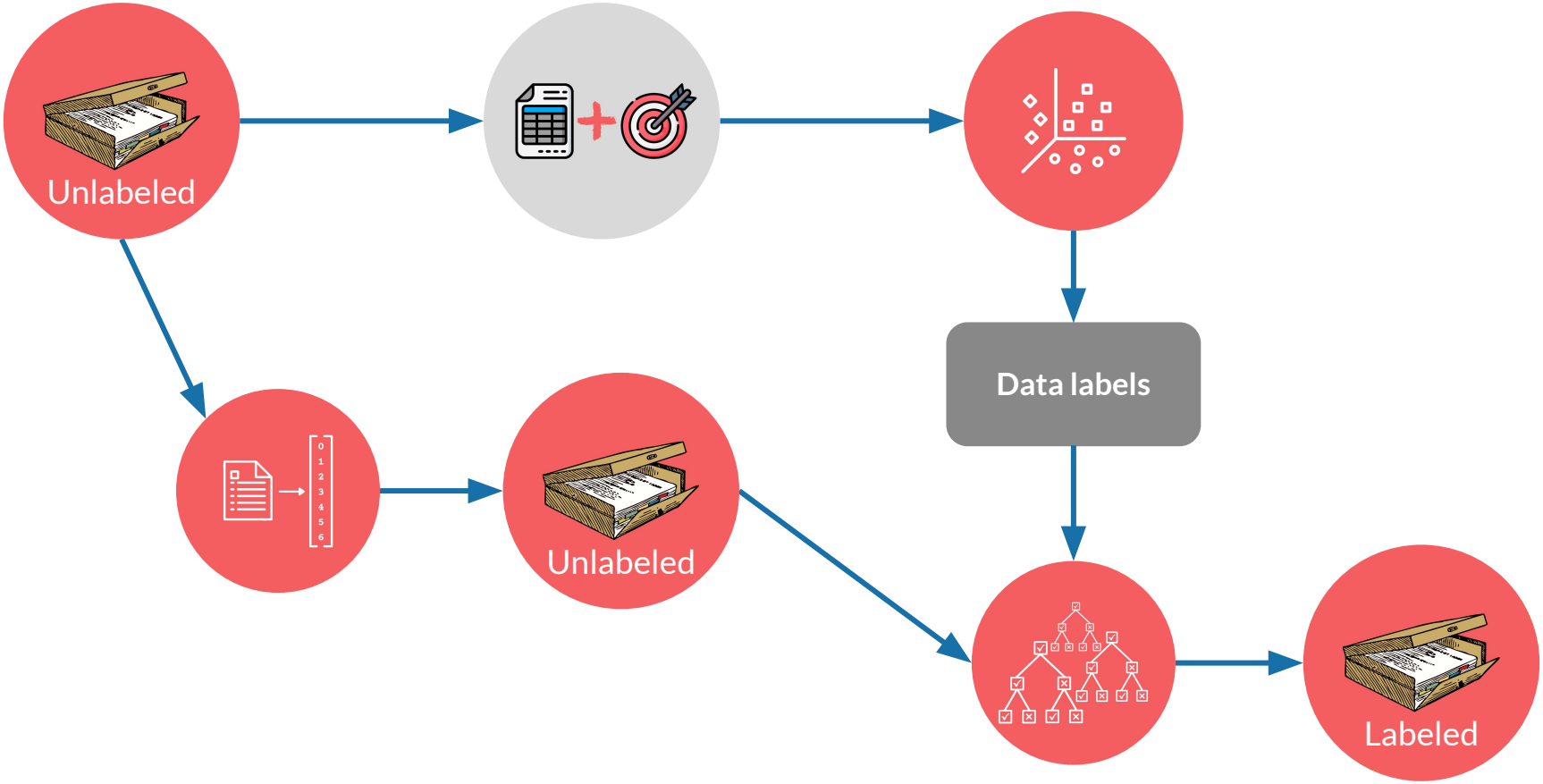
1. Generalizable


Unlabeled
Data labels
Unlabeled
Labeled

2. Dynamic


Unlabeled
Vocabulary is built afresh each time the corpus changes sealing in word order and document context.
Data labels
Unlabeled
Labeled

3. Interpretable



Samples from the corpus and overall distribution of positive and negative sentiment can be used to determine nature of groupings.

4. Relevant

Benefits of #KOT style modeling

Meet Kevin.

Photo c/o Reuters

Benefits of #KOT style modeling

- Better curation of Kenyan written, audio and audiovisual content
- More accurate results from automated media monitoring
- Integration of audience feedback into media buying

Photo c/o Reuters

Benefits of #KOT style modeling

Meet the team revolutionizing business.

Photo c/o Safaricom on Flickr

Benefits of #KOT style modeling

- Better automated analysis of customer feedback on written, audio and audiovisual format
- Integration of stakeholder feedback into product development, pricing and go-to-market strategy

Photo c/o Safaricom on Flickr

Benefits of #KOT style modeling

Meet Moses.

Photo c/o Sirdart on CitizenTV

Benefits of #KOT style modeling

- Integration of stakeholder feedback into monitoring and evaluation
- Tracking evolution of Sheng
- Production of relevant media that resonates with local communities

Photo c/o Sirdart on CitizenTV

Some benefits of including informal language for language analysis may include:

A. tracking changes in slang / sheng
B. better automated customer service
C. improved understanding of community needs
D. all of the above

Some benefits of including informal language for language analysis may include:

D. all of the above

Conclusion

We need either more data or more time to grow a Kenyan corpus.


50M
Kenyan citizens
8K
English speaking users of #KOT in 2020-2021

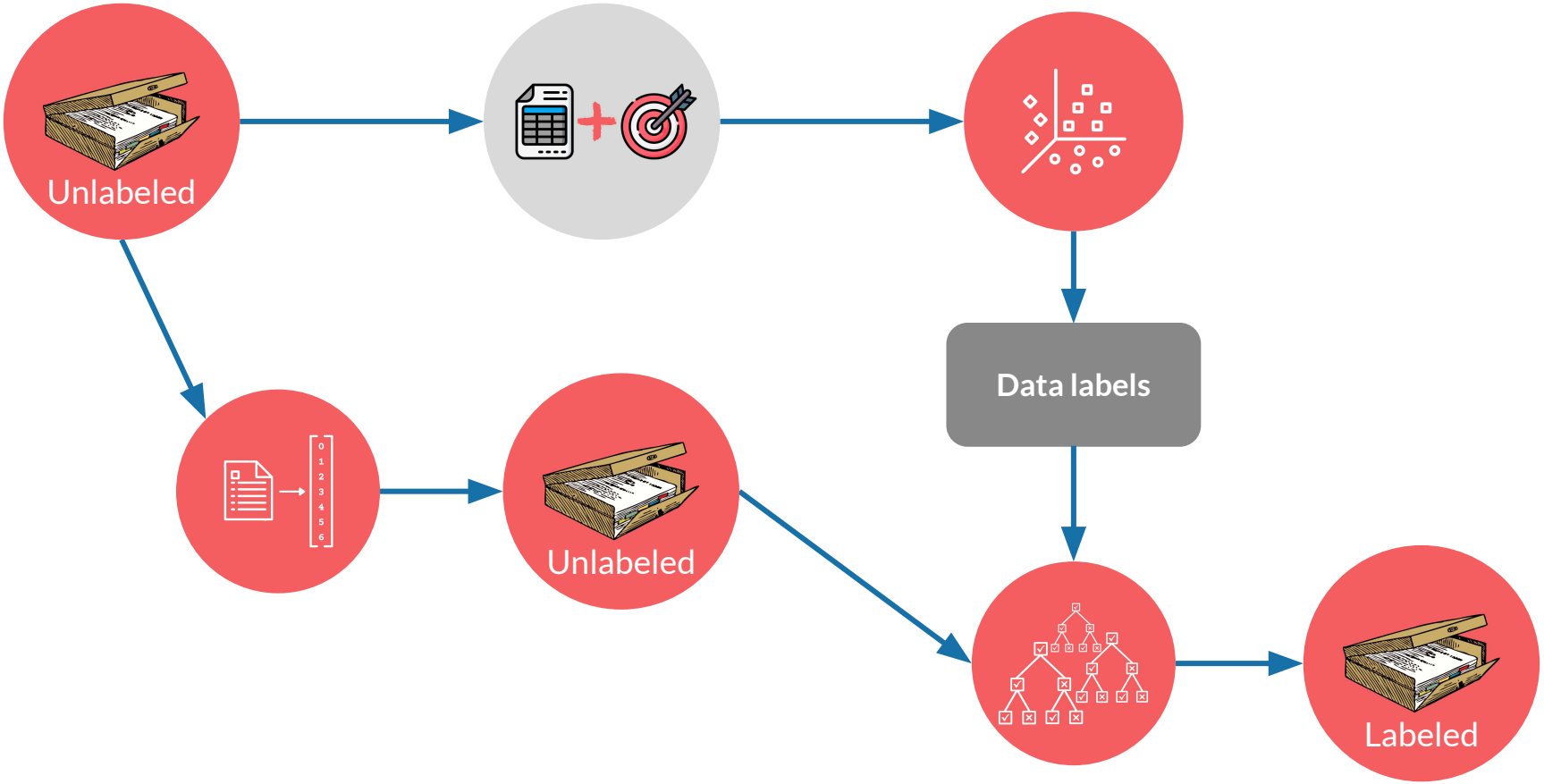
Possible Additional Datasets

- Kenyan Facebook Groups
- Kenyans on Linkedin
- Kenyans on Instagram
- Kenyan content on YouTube

What next?

#KOT Style Modelling

- Using this style on other multilingual, code-switching cultures
- Using this style with audio content
- Using this style with video content


Unlabeled
Data labels
Unlabeled
Labeled

Kenyans on Twitter

- Investigating why many #KOT tweets have neutral sentiment
- Predict the political tweets without using hashtags
- Predict which tweets are ads or spam

We need a solid framework to better scope and implement NLP projects...

...so I created the NLP Toolbox here and used it in this talk.

What is the NLP toolbox?

The NLP toolbox is a collection of concepts, tools and ideas available for building applications that can handle real-world challenges around understanding content from all over the world.

It can be used for any NLP task such as sentiment analysis, translation, transcription, audio synthesis etc.

Resources & Contacts

1. https://github.com/CeeThinwa/Delta-Analytics-2021-CT-Project

2. NLP beginner resources
    a. Deep Learning for NLP resources: https://github.com/andrewt3000/DL4NLP
    b. Natural Language Processing in Python: https://youtu.be/xvqsFTUsOmc
    c. Natural Language Process in 10 minutes: https://youtu.be/6I-Alfkr5K4

More questions?

Email: ceethinwa@gmail.com
LinkedIn: https://ke.linkedin.com/in/cynthiathinwa

Acknowledgements

To my incredible coach Ugaso,

A BIG THANK YOU!

To the incredible Delta Analytics organisers and facilitators,

A BIG THANK YOU!

To my fellow 2021 Delta Analytics Teaching Fellows,

A BIG THANK YOU!

To you, our audience,

A BIG THANK YOU!

To my family, friends and God,

A BIG THANK YOU!

Any Questions?