

TANZANIA DATA LAB (DLAB)



GROUP 4: SWAHILI SOCIAL MEDIA SENTIMENT ANALYSIS PROJECT

Contents

INTRODUCTION	3
1.1 BACKGROUND	3
1.2 PROBLEM	3
1.3 OBJECTIVE	3
1.3.1 SPECIFIC OBJECTIVES	3
1.4 TARGET.....	4
1.5 PROJECT IMPACT.....	4
1.6 TEAM MEMBERS	4
DATA	5
2.1 DATA SOURCES	5
2.2 DATA CLEANING	6

List of figures

Figure 1:Team members.....	4
Figure 2: The data shape.....	5
Figure 3: The dataset datatypes.....	6
Figure 4: Cleaning the dataset.....	6

INTRODUCTION

1.1 BACKGROUND

The project is based on a dataset collected through Twitter social media across East African countries that speak Swahili language. The dataset was collected in countries from East Africa (Tanzania, DRC, Kenya and Uganda) .

The project is aimed to analyze social conversations online and determine deeper context as they apply to a topic, brand or theme.

1.2 PROBLEM

Analysis of social conversations online has been challenged because most Swahili speakers tend to express themselves in their own local dialect. Conversations online concerning various topics, themes or matters pertaining the society including business have been somewhat meaningless to non-Swahili speakers who tend to be directly or indirectly concerned with the discussed aspects.

A relevant group in this context could be banking, brands, insurance companies, social media influencers or social media owners who might have difficulty understanding and interpreting product audience and their reactions.

1.3 OBJECTIVE

The objective of this project is to determine whether a Swahili sentence is of positive, negative, or neutral sentiment.

1.3.1 SPECIFIC OBJECTIVES

- To perform data cleaning so that it can be easily manipulated.
- To build a model that can predict the polarity of new Swahili statements particularly tweets.
- The solution to be built could be used in automatic speech recognition(ASR) technology services.
- The solution could also be used by banking, insurance companies, or social media influencers to better understand and interpret a product's audience and their reactions.

1.4 TARGET

This solution could target banking, insurance companies, social media influencers, social media owners and brands.

1.5 PROJECT IMPACT

The positivity, neutrality and negativity of Swahili statements written by users could be a source of information which would impact in decision making hence development of businesses at large.

1.6 TEAM MEMBERS

NAME	COLLEGE	YEAR OF STUDY
NAJMA Y MAJID	COICT, UDSM	III
LUCAS G MAHIKIMBA	DIT	I
SAMWEL J KAHUNGWA	COICT, UDSM	III

Figure 1:Team members

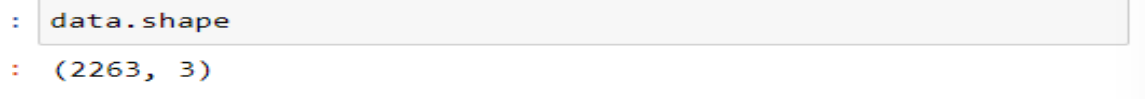
DATA

2.1 DATA SOURCES

The dataset is from Zindi Africa platform. Zindi Africa with EA ambassadors gathered Swahili contents (tweets) from Twitter that express labels about statement polarity. The data was preprocessed by removing links, emoji symbols, and punctuations.

The collected tweets were manually annotated using an overall polarity: positive (1), negative (-1) and neutral (0).

The dataset has 2263 rows and 3 columns.



```
: data.shape
: (2263, 3)
```

Figure 2: The data shape

The columns are ID, Tweets and Labels where:

ID - This is the unique ID of a unique Swahili tweet.

Tweets - This is the content of a unique tweet.

Labels - This is a sentiment of a particular tweet (positive (1), negative (-1) and neutral (0)).

The dataset can be obtained from [Google NLP Hack Series: Swahili Social Media Sentiment Analysis Challenge - Zindi](#)

2.2 DATA CLEANING

➤ Checking the data types in the dataset

- The dataset had int64 and object data types

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2263 entries, 0 to 2262
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype  
---  --
 0   ID          2263 non-null   int64  
 1   Tweets      2263 non-null   object  
 2   Labels      2263 non-null   int64  
dtypes: int64(2), object(1)
```

Figure 3: The dataset datatypes

➤ Cleaning the features of the dataset

The column “Tweets” in the dataset contains Swahili statements. We cleaned the statements by converting capital letters to small letters, removing numbers, links and punctuations so as to make the words in the statements to be trained well in the model.

```
def text_cleaning(text):
    # Clean the text
    text = re.sub("[^a-zA-Z]", " ", text.lower()) #convert capital letters to small letters
    text = re.sub(r"\s", " ", text) #remove punctuations
    text = re.sub(r'http\S+', ' link ', text) #remove links
    text = re.sub(r'\b\d+(?!\.\d+)?\s+', '', text) # remove numbers
    return(text)
```

```
data["cleaned_tweets"] = data["Tweets"].apply(text_cleaning)
```

```
data.head(7)
```

	Tweets	Labels	cleaned_tweets
0	So chuga si tunakutana kesho kwenye Nyamachoma...	0	so chuga si tunakutana kesho kwenye nyamachoma...
1	Asante sana watu wa Sirari jimbo la Tarime ...	1	asante sana watu wa sirari jimbo la tarime ...
2	Leo nimepata kitambulisho changu cha taifa ...	1	leo nimepata kitambulisho changu cha taifa ...
3	Mgema akisifiwa tembo hilitia maji	0	mgema akisifiwa tembo hilitia maji
4	Ee Mwenyezi Mungu Msamehe na Umrehemu na Umuaf...	1	ee mwenyezi mungu msamehe na umrehemu na umuaf...
5	Kama wewe ni mfanyabiashara au mfanyakazi u...	1	kama wewe ni mfanyabiashara au mfanyakazi u...
6	Asante na tafadhali subiri tutawasiliana nawe ...	1	asante na tafadhali subiri tutawasiliana nawe ...

Figure 4: Cleaning the dataset