جهـــاز التخطيـــط والإحصـــاء
Planning and Statistics Authority
State of Qatar • دولـــة قطـــر

Regional Workshop on the Changing Role of Official Statistics in the State of Qatar:
Why Data Culture Matters

21st -22nd September 2022, Doha, Qatar
Pullman Hotel, Al Thuraya Ballroom - West Bay

ورشـة العمل الإقليمية حول الـدور المتغير للإحصاءات الرسـمية فـي دولة قطر:
ثقافة البيانات مهمة

٢١-٢٢ سبتمبر ٢٠٢٢، الدوحة، قطر
فندق بولمان الدوحة، قاعة الثريا - الخليج الغربي

UNITED NATIONS DEVELOPMENT PROGRAMME

UNDP UKRAINE COUNTRY OFFICE

USING ML/NLP IN TRACKING PROGRESS TOWARDS SDGS

INTRODCUTION

Social Listening (SL) is a technique to monitor and collect useful information form social media. It employs social media to collect the pulse of the crowds and find the trends of discussions at the grassroots of the community. Unlike traditional methods of data collection (that rely on predesigned questionnaires), answers are collected directly from social media portals and analyzed to discover people's interests.

In SL, keywords and mentions are used to filter collected data and highlight the results that worth further investigation. Instead of reading astronomic number of social media posts, SL filters out unwanted posts, and consequently reduces text corpus to a suitable size for further analysis. Despite being cheap and fast, using keywords and mentions is not the best option, where keywords and mentions come in many different forms and shapes that – sometimes – confuses machines and makes them – in many cases – useless.

We propose to use Natural Language Processing (NLP) to map social media posts to SDGs where we use NLP to "semantically search" for posts of relevance to SDGs. In this approach, *we do not search for keywords or mentions* but rather for the semantics and meanings behind those keywords and mentions. Those semantics are harvested from pretrained generic language models (as we will see next).

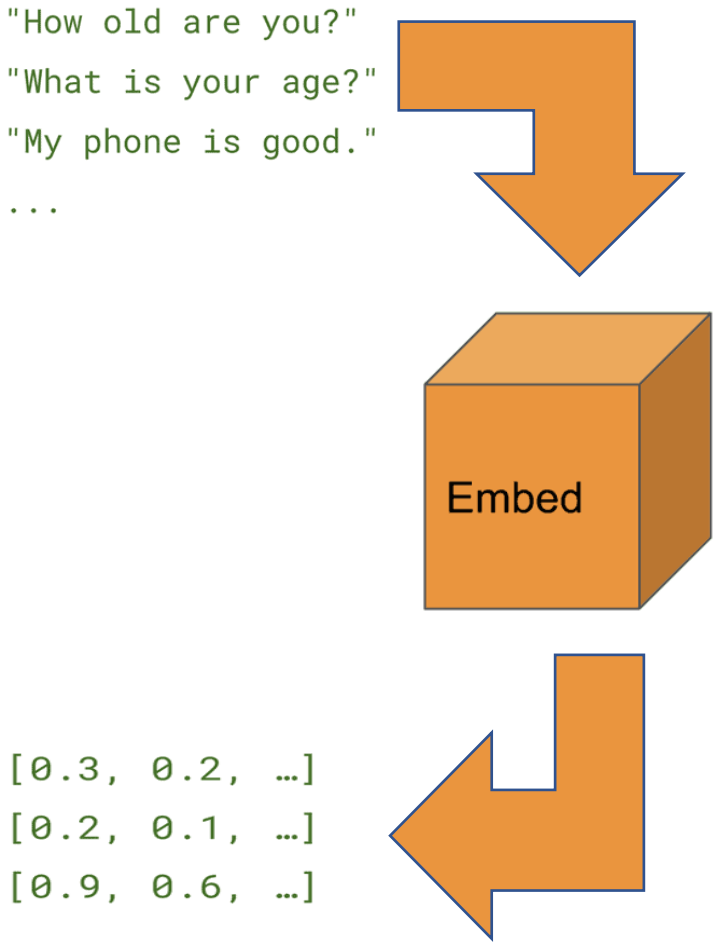
METHODOLOGY

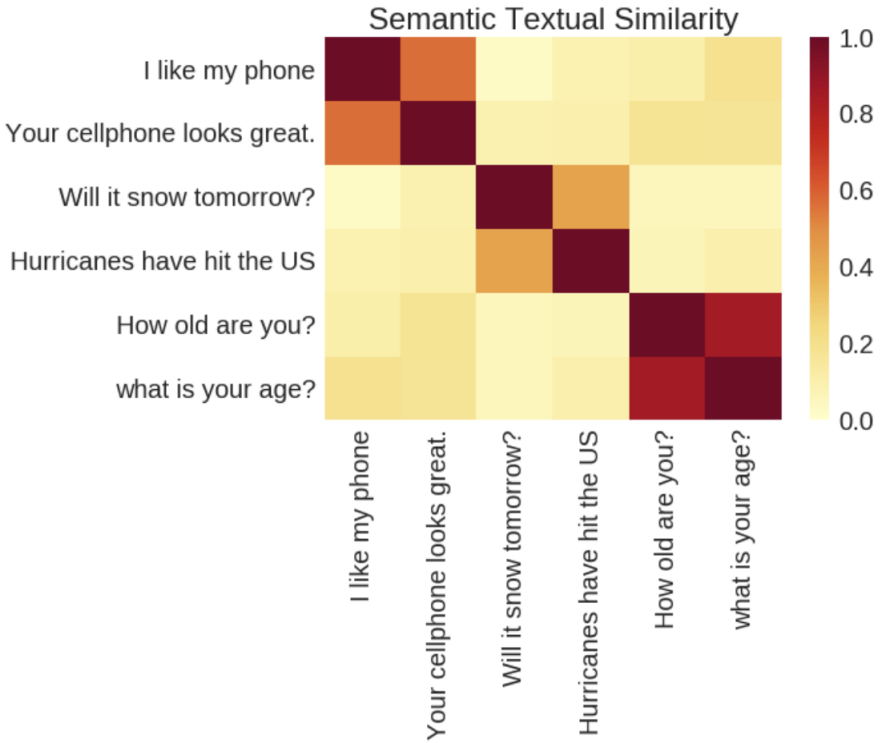
Using Machine Learning (ML) in NLP tasks is a quite known and widely used approach in recent NLP research. ML is used in recent studies to train language models and use them in different NLP tasks. While many techniques can be used in ML models training, we focus on Artificial Neural Networks (ANN).

Using ANN to create word, sentence, and document representations has gained growing interest from research community in the past decade. From Word2Vec and fastText to BERT and GPT, scientists were improving the performance of computer-aided language models in addressing day-to-day challenges and automating language processing tasks to replace – the slow and expensive – human agents.

Most of the recent language models, produced by NLP community, take the form of sentence embeddings. In those embeddings, sentences are represented by vectors of weights/numbers where each sentence in training corpus take a specific composite of numbers. Similar sentences are represented by similar vectors, while dissimilar sentences are represented by dissimilar vectors. Geometric distance is used at later stage to distinguish between similar and dissimilar sentences.

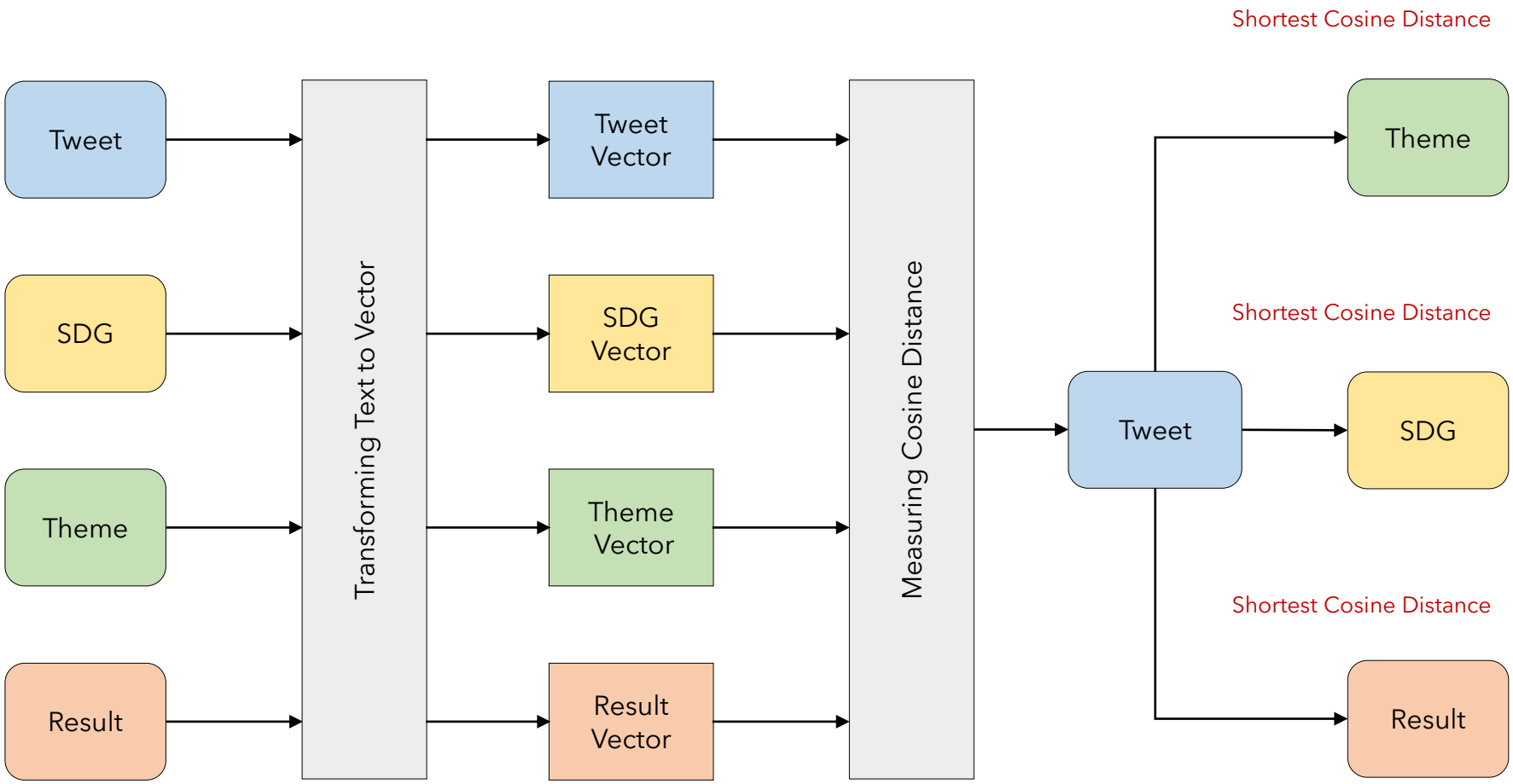
ILLUSTRATIONS





Source: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46808.pdf

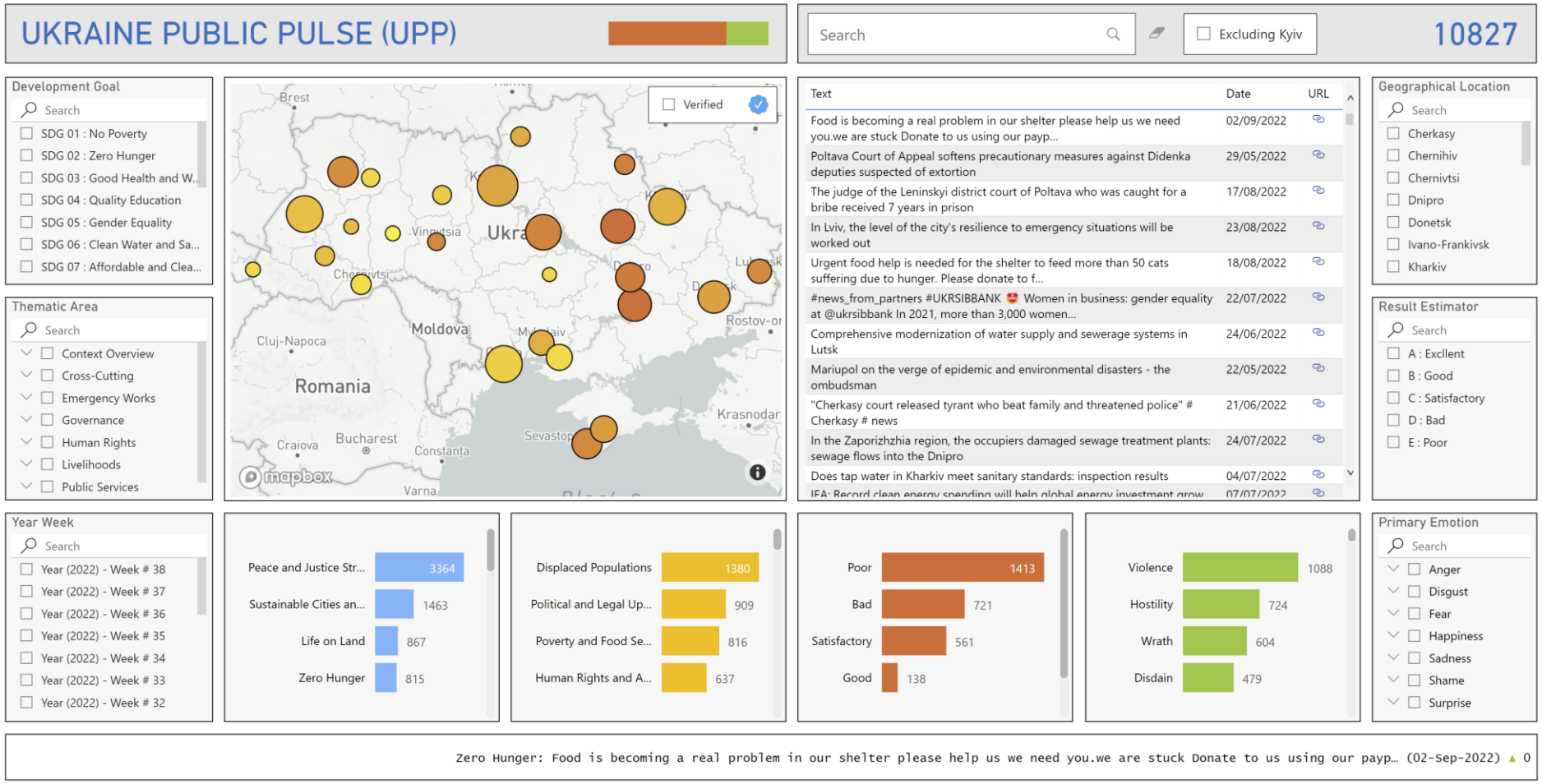
OUR MODEL


Tweet
SDG
Theme
Result
Transforming Text to Vector
Tweet Vector
SDG Vector
Theme Vector
Result Vector
Measuring Cosine Distance
Tweet
Shortest Cosine Distance
Theme
Shortest Cosine Distance
SDG
Shortest Cosine Distance
Result

RESULTS

Tweet - INPUT	**SDG - OUTPUT**	**Similarity**
Life in Odessa → Unhealthy landfill must be eliminated on irrigation fields	SDG 15: Life on Land	0.415104
"Relevant for citizens under martial law" #Cherkasy #news	SDG 16: Peace and Justice Strong Institutions	0.313918
Weather forecast for September 8: air and water temperatures equalized on the Black Sea coast ...	SDG 13: Climate Action	0.303138
The High Council of Justice supported the selection of judges of the Constitutional Court with the help of foreign experts 👉...	SDG 16: Peace and Justice Strong Institutions	0.296112
My favorite city is Donetsk! My soul and heart is always with you! You are strong! You can do it!🤍	SDG 11: Sustainable Cities and Communities	0.284486
The scandalous developer of Dignity Square tried to "reclaim" his MAF through the court #Lviv	SDG 16: Peace and Justice Strong Institutions	0.282028
Almost all of Ukraine's grain is sent not to developing and poorer countries, but to EU - #Putin #Europe #Russia	SDG 01: No Poverty	0.274340
#NEFCO News #Infrastructure #Ukraine NEFCO will allocate 50 million euros to restore critical infrastructure…	SDG 09: Industry, Innovation and Infrastructure	0.260868
The commander of the battalion Troop "Azov" (Dnepr) about the intentions to resume blackmail with a court of…	SDG 16: Peace and Justice Strong Institutions	0.260036

RESULTS

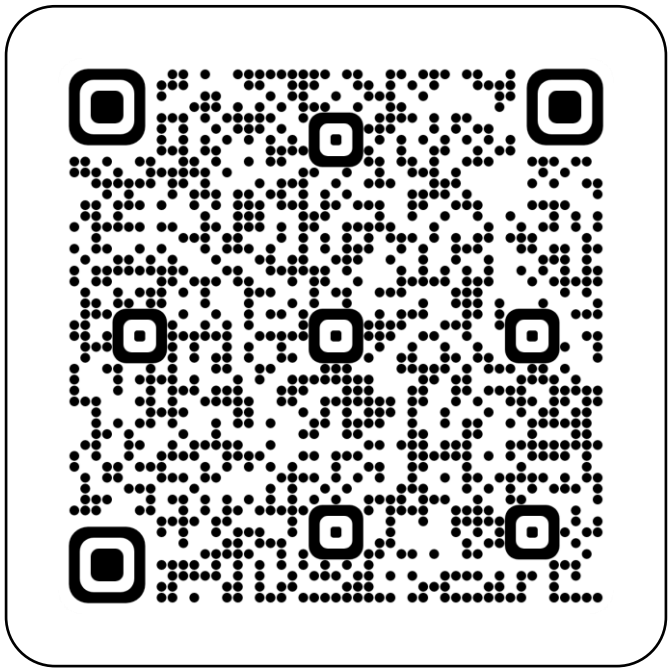

UKRAINE PUBLIC PULSE (UPP)
Search
Excluding Kyiv
10827
Development Goal
Search
SDG 01 : No Poverty
SDG 02 : Zero Hunger
SDG 03 : Good Health and W...
SDG 04 : Quality Education
SDG 05 : Gender Equality
SDG 06 : Clean Water and Sa...
SDG 07 : Affordable and Clea...
Thematic Area
Search
Context Overview
Cross-Cutting
Emergency Works
Governance
Human Rights
Livelihoods
Public Services
Verified
Brest
Vinnytsia
Chernivtsi
Moldova
Cluj-Napoca
Romania
Craiova
Bucharest
Constanța
Varna
Sevastopol
Krasnodar
Rostov-on
mapbox
Text
Date
URL
Food is becoming a real problem in our shelter please help us we need you.we are stuck Donate to us using our payp... 02/09/2022
Poltava Court of Appeal softens precautionary measures against Didenka deputies suspected of extortion 29/05/2022
The judge of the Leninskyi district court of Poltava who was caught for a bribe received 7 years in prison 17/08/2022
In Lviv, the level of the city's resilience to emergency situations will be worked out 23/08/2022
Urgent food help is needed for the shelter to feed more than 50 cats suffering due to hunger. Please donate to f... 18/08/2022
#news_from_partners #UKRSIBBANK 😍 Women in business: gender equality at @ukrsibbank In 2021, more than 3,000 women... 22/07/2022
Comprehensive modernization of water supply and sewerage systems in Lutsk 24/06/2022
Mariupol on the verge of epidemic and environmental disasters - the ombudsman 22/05/2022
"Cherkasy court released tyrant who beat family and threatened police" # Cherkasy # news 21/06/2022
In the Zaporizhzhia region, the occupiers damaged sewage treatment plants: sewage flows into the Dnipro 24/07/2022
Does tap water in Kharkiv meet sanitary standards: inspection results 04/07/2022
Geographical Location
Search
Cherkasy
Chernihiv
Chernivtsi
Dnipro
Donetsk
Ivano-Frankivsk
Kharkiv
Result Estimator
Search
A : Exclent
B : Good
C : Satisfactory
D : Bad
E : Poor
Year Week
Search
Year (2022) - Week # 38
Year (2022) - Week # 37
Year (2022) - Week # 36
Year (2022) - Week # 35
Year (2022) - Week # 34
Year (2022) - Week # 33
Year (2022) - Week # 32
Peace and Justice Str... 3364
Sustainable Cities an... 1463
Life on Land 867
Zero Hunger 815
Displaced Populations 1380
Political and Legal Up... 909
Poverty and Food Se... 816
Human Rights and A... 637
Poor 1413
Bad 721
Satisfactory 561
Good 138
Violence 1088
Hostility 724
Wrath 604
Disdain 479
Primary Emotion
Search
Anger
Disgust
Fear
Happiness
Sadness
Shame
Surprise
Zero Hunger: Food is becoming a real problem in our shelter please help us we need you.we are stuck Donate to us using our payp… (02-Sep-2022) ▲ 0

CHALLENGES

1. Absence of labelled text corpora. We don't have labelled SDGs text corpus. Hence, we can't train models on SDGs text classification or other NLP tasks. Therefore, we used pretrained language model to address this issue. Having labelled text corpus, in future, will make it possible to train a custom model to classify text into SDGs.
2. Different descriptions yield different semantic similarities. The question is: what are the right words and sentences to describe each SDG? There is no easy answer to this question though we noticed that shorter descriptions tend to yield better results.
3. Deciding on similarity threshold is a challenge. Sometime a low similarity between a Tweet and SDG is enough to show a valid correlation. In other times, higher similarity is not enough to map a Tweet to another SDG. Using dynamic similarity might help in addressing this issue. This assumption can be tested once the model is manually evaluated (Next Slide, Rec. 3).
4. Using Sentence Encoding in text classification is fairly new technique. References and prior work in this domain are limited. Hence, most of the experiments we conduct have no benchmarks to compare to and most of the work has to be done from scratch. This will change in future.

THE WAYFORWARD

1. Text is classified into 17 SDG, better to have classification go down to Targets and Indicators level. Instead of classifying text into SDGs only, we may extend it to cover the whole set of 169 targets and 232 indicators. More granular and better classification can be achieved.
2. We used the locations of the major 26 cities in Ukraine, we may extend this to lower administrative level. We have about a 1000 Ukrainian settlement in our database that can be used all in this process.
3. Manual evaluation of results to improve classification accuracy, precision, and recall. Use evaluated model to produce a labelled text corpus to be used in training a new model on SDG classification.
4. Fusion of textual resources with other types of data, such as satellite images, might be a good idea to be tested and evaluated. Integrate Computer Vision (CV) with NLP, in future, to infer new knowledge by training ML models on both texts and images to predict progress towards SDGs.
5. Integrate the predictions of this model in future policies related to SDGs. This model can be used to advise policy makers on a) geographic and thematic areas that need improvements, b) where the performance is below expectations, and c) where to devote more efforts and resources.
6. Extend the model to other countries and compare it results to Human Development Reports (HDR). Such extension helps in finding linguistic patterns in places where SDGs are successfully implemented.



https://tinyurl.com/uapua

THANK YOU