https://doi.org/10.2352/EI.2023.35.3.MOBMU-356

Practical OSINT Investigation -– Similarity Calculation using Reddit User Profile Data

Valeriya Vishnevskaya[1]***, Klaus Schwarz***[1,3]***, Reiner Creutzburg***[1,2]

[1] ***SRH Berlin University of Applied Sciences, Berlin School of Technology, Ernst-Reuter-Platz 10, D-10587 Berlin, Germany Email: lera071992@gmail.com, klaus.schwarz@srh.de, reiner.creutzburg@srh.de***

[2] ***Technische Hochschule Brandenburg, Department of Informatics and Media, IT- and Media Forensics Lab, Magdeburger Str. 50, D-14770 Brandenburg, Germany, Email: creutzburg@th-brandenburg.de***

[3] ***University of Granada, Faculty of Economics and Business, P.◦ de Cartuja, 7, ES-18011 Granada, Spain***

Keywords: Open Source Intelligence, OSINT, SOCMINT, Reddit, Cybersecurity, Cyber Security, OSINT investigation, cybersecurity training

Abstract

This paper presents a practical Open Source Intelligence (OSINT) use case for user similarity measurements using open profile data from the Reddit social network. This PoC work combines the open data from Reddit and part of the state-of-the-art BERT model. Using the PRAW Python library, the project fetches comments and posts of users. Then these texts are converted into a feature vector representation of all user posts and comments. The main idea here is to create a comparable user's pair similarity score based on their comments and posts. For example, if we fix one user and calculate the scores of all mutual pairs with other users, we will produce a total order on the set of all mutual pairs with that user. This total order can be described as a degree of written similarity with this chosen user. A set of "similar" users for one particular user can be used to recommend to the user interesting for him people. The similarity score also has a "transitive property": if $user_1$ is "similar" to $user_2$ and $user_2$ is similar to $user_3$ then inner properties of our model guarantee that $user_1$ and $user_3$ are pretty "similar" too. This way, this score can be used to cluster users into sets of "similar" users. It could be used in some recommendation algorithms or tune already existing algorithms to consider a cluster's peculiarities. Also, we can extend our model and calculate feature vectors for subreddits. In that way, we can find similar to the user's subreddits and recommend them to him.

OSINT Basics

Open Source Intelligence (OSINT) is an information-gathering, collecting, and analyzing process using open and public resources. Since the end of the Cold War, the world is becoming more global and open, and the invention of WWW brings technology to small places in the countryside. This vital transformation brings huge computerization and relevant benefits to society. However, at the same time, it triggers different kinds of risks, for example, oppressive regimes, cybercriminals, terrorist groups, and all using the World Wide Web for conducting their crimes. By 2019 money lost from those crimes will cost businesses more than 2 trillion dollars. That critical risk encourages governments to invest in research and development of open source intelligence (OSINT) techniques and tools to handle future cybersecurity challenges and corner cases of that. OSINT operates with information that is free and available in public resources. Different organizations from different sectors provide OSINT services - Open Source Centers, government, BBC Monitoring, and Private Sector organizations. BBC Monitoring is a department within the British Broadcasting Corporation (BBC) that does some monitoring for foreign media worldwide. It was first established in 1939 and had offices in different countries around the globe. Monitoring includes Internet, radio, TV broadcasting, emerging trends, and print media through 150 countries in more than 70 languages. The WWW can be consistently divided into the deep web and surface web. Most information is easily accessible using a web search engine because the search engine indexes the content. This part of the Internet is called the surface web. In contrast, the deep web can only be accessed by direct connection using the Unique Resource Locator (URL) or IP address because a search engine does not index its content. Some of these destinations require registration and sometimes also payment. Another part of the Internet is the named darknet, which requires specific software, small peer-to-peer networks, configuration, and authorization to access. The dark web is situated on top of the darknet. In addition, some sites are referred to as hidden services, which are only reachable using the darknet. Thompson lists different darknet technologies – FreeNet, TOR, and I2P which can be named overlay networks. An existing network uses overlay networks to create a new layer on top of it.

All publicly accessible sources of information are held in OSINT. That information can be searched offline and online, including in the following places:

1. The Internet: video-sharing sites like YouTube.com, social networking sites, blogs, Whois records of registered domain names, forums, wikis, digital files metadata, dark web resources, IP addresses, people search engines, geolocation data, and whatever could be easily found online.
2. Traditional mass media: newspapers, books, radio, TV,

books, and magazines.
3. Academic publications, specialized journals, conference proceedings, dissertations, annual reports, company profiles, employee profiles, company news, CVs.
4. Videos i Photos and including metadata
5. Geospatial information: commercial image products and maps

OSINT intelligence used by business organizations for other non-financial purposes:

1. Security vulnerabilities of business networks and exposure of confidential information are reasons for future cyber threats, and companies are trying to fight against that data leakage.
2. To create their threat intelligence strategies through analyzing OSINT sources from outside and inside a business and combining information with other information to accomplish an effective cyber-risk management policy that helps them protect the reputation and customer base financial interests.

OSINT is significantly helpful for companies specializing in the defense industry, who must be aware of the potential offenses from customers to create appropriate equipment.

OSINT is used extensively by hackers and penetration testers to gather intelligence about a specific target online. It is also considered a valuable tool for conducting social engineering attacks. The first phase of any penetration testing methodology begins with reconnaissance (in other words, with OSINT).

Figure 1. *Phases of penetration testing*

Companies pay penetration testers to break into internal networks to show where weaknesses lie and how to keep outsiders out. This differs from black hat hackers who exploit these vulnerabilities to gain unauthorized access to confidential data; however, both use the same reconnaissance techniques and tools to achieve their work.

Reddit

Reddit is the most prominent American social news aggregator, discussion platform, and content rating network website. Users who registered there named Redditors, users allowed to submit different types of content such as text posts, videos, images, links, and images. The primary mechanism of the application works like that - all posts can be voted up or down by other members/users of the Reddit social network. Post here organized by subject labeled topic use-creation - 'subreddits' or 'communities .' Submissions with more upvotes appear towards the top of their subreddit and, if they receive enough upvotes, ultimately, on the site's front page - Reddit administrators moderate communities. The moderators' stack runs by community-specific responsible moderators who are not working for Reddit as employees. In 2022, Reddit ranked the 9th-most-visited website in the world and sixth place in the same list of most-visited in the USA, according to Semrush. The user base here is approximately 42 to 49.3% from the USA, from the UK 7.9 to 8.2%, and from Canada, between 5.2 and 7.8%. On a regular daily basis, Reddit uses 22% of US adults between 18 to 29 years and 14% of U.S. adults aged 30 to 49 years. Let us look at more detailed Reddit statistics information:

1. Reddit has 52 million daily active users;
2. Reddit has over 430 million monthly active users;
3. Reddit has raised a total of $1.3 billion in funding.
4. Reddit is worth $10 billion.
5. 52 million daily active users access Reddit.
6. 25% of US adults use Reddit.
7. Reddit was ranked the 9th most popular social media app in the US.
8. 48% of Reddit visitors are in the US.

Reddit recently switched from disclosing monthly users to focusing on daily active users. However, there is plenty of historical data on how many users visit the platform at least monthly. As of 2019, Reddit had 430 million monthly active users (MAUs), a growth of 30.3% compared to a year earlier.

Year	Number of monthly active users (MAUs)
2012	**46 million**
2013	**90 million**
2014	**174 million**
2015	**199 million**
2017	**250 million**
2018	**330 million**
2019	**430 million**

Figure 2. *Breakdown of Reddit MAU growth last year [6]*

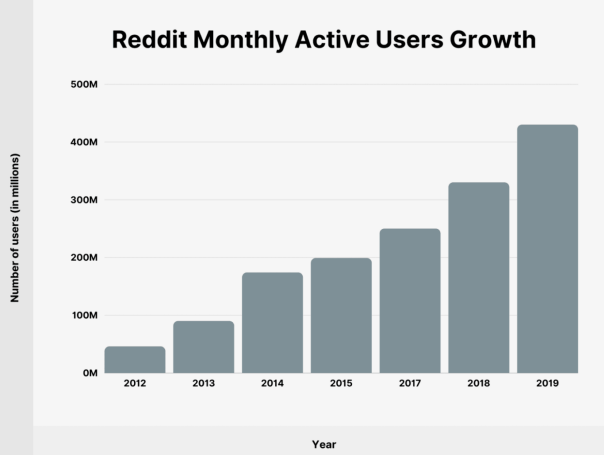

Figure 3. *Reddit monthly active users growth [6]*

As of August 2021, 48% of Reddit users are in the US, followed by the UK, Canada, Australia, and Germany.

Reddit founders were students from the University of Virginia. They were friends and roommates and started a business together in 2005. The founders are Alexis Ohanian and Steve Huffman, Aaron Swartz. Reddit raised $50 million in a funding round led by Sam Altman, including investors Peter Thiel, Marc

Country	Distribution of Reddit users
United States	**47.82%**
UK	**7.6%**
Canada	**7.45%**
Australia	**3.89%**
Germany	**3.37%**

Figure 4. *Countries with the largest Reddit user bases [6]*

Andreessen, Ron Conway, Jared Leto, and Snoop Dogg in October 2014. Their investment valued the company at $500 million then. In July 2017, Reddit raised $200 million for a $1.8 billion valuation, with Advance Publications remaining the majority stakeholder. In February 2019, Tencent's $300 million funding round brought the company's valuation to $3 billion. Finally, in August 2021, a $700 million funding round led by Fidelity Investments raised that valuation to over $10 billion. In January 2021, Reddit had its highest monthly downloads. The 6.6 million global downloads represented a 2x increase over the previous year.

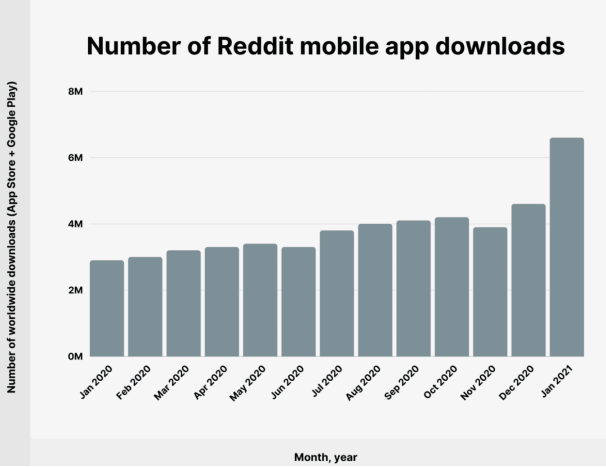

Figure 5. *Number of Reddit mobile app downloads [6]*

A Business Insider Intelligence analysis found that 62% of US Reddit users agree that the platform protects their privacy and data. Additionally, 15% of Reddit users "strongly agree" that Reddit protects their privacy and data.

Common themes among the top 10 subreddits include humor, gaming, news, science, and Reddit. Reddit is smaller than other social media players, like Instagram, Facebook, and Instagram, but has grown dramatically. Events surrounding the Wall Street Bets subreddit have given an additional boost to Reddit's popularity.

OSINT and Reddit

Many tools help investigate and collect Reddit information using available public information. OSINT framework focused on gathering information from free tools or resources. The intention is to help people find free OSINT resources. Some of the sites included might require registration or offer more data in the paid version, but one should be able to get at least a portion of the available information for no cost. This paper created the framework from an information security point of view. Since then, the response from other fields and disciplines has been incredible. Let us look at tools that allow us to collect information from Reddit.

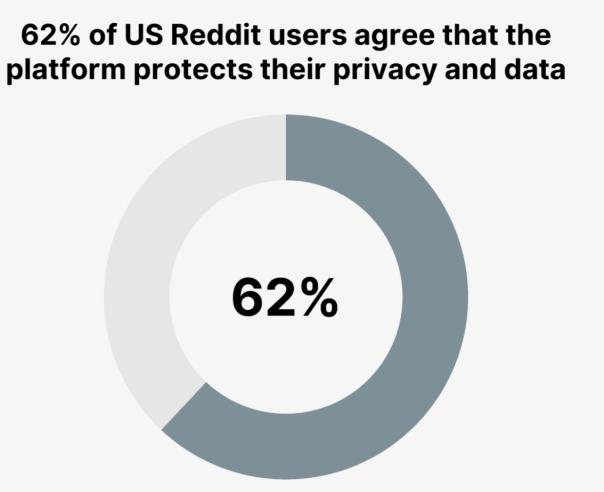

Figure 6. *Reddit's digital trust & privacy reputation [6]*

Useful tools to extract some open information from Reddit:

1. Pushshift search - Search engine and analytics tracker for Reddit
2. Redective - Gives information on Reddit users. Search by username
3. Karma - Decay Reverse image search. Shows whether the image has been posted to Reddit
4. r/whatisthisthing - Subreddit crowdsourcing identification of objects submitted by users

BERT Model

This project uses a pre-trained BERT (Bidirectional Encoder Representation of Transformers) deep bidirectional language representation model. It was trained from unlabeled trillions of pages of text from Wikipedia and is very helpful in understanding language with only text content; suitable for finding solutions for text classification, for example, language inference and question-answering tasks without specific settings. BERT is empirically powerful and conceptually simple. BERT is in use at Google Search Engine.

Attention

Traditionally, what you can do if you are interested in NLP, is if you have a language task, 'The cat eats the mouse,' and you like to translate it to another language - German. You need to encode that text to representations and decode it again. The sentence needs to be transformed to the vector, which needs to be translated somehow to the target language - traditional seq to seq task. LST Matching Networks are very popular for that task. You go over the source sentence one by one. Take the word 'The' and encode it to the vector and use the neural network to turn the vector to the hidden state Ho, after we take the second token cat', to represent it to the vector too and put it through the same function Hstart hidden state here to predict another hidden state H1. We do the same with all words in sentences. We would use the last hidden

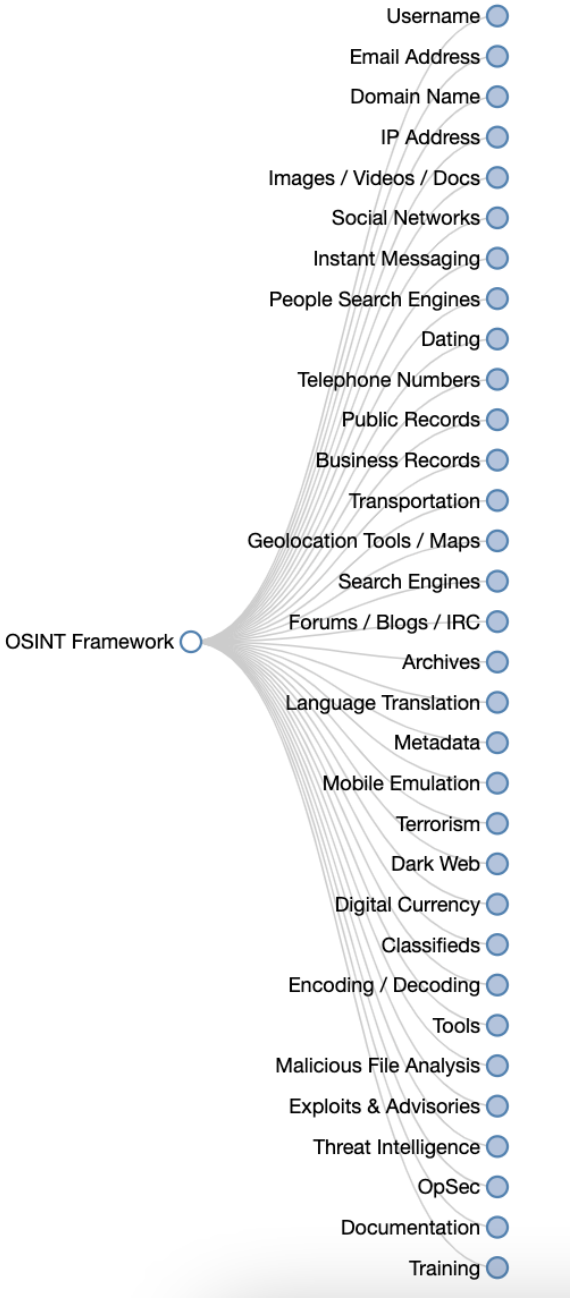

Figure 7. *OSINT framework [11]*

state H4 in the same fashion to put it to the decoder (Die); the outcome will be the word for the next hidden layer H5 and H5 will go again to the decoder 'Katze'. It takes the current input and the last hidden state and computes the new one. In the decoder case, it takes the hidden state and the previous word you output to fit the decoder - the next word. The incode uses the sentence's meaning and the last word you need for grammar; for example, the next word will be based on that. Attention is the mechanism here to increase performance here. If we look at the decoder here for the word 'eats' - 'Essen'. The only information the system has is the last word and hidden state. If we look at what the actual word in the sentence should be output there, it is 'eats'. Output averaged hidden state can be used as vectorized text representation.

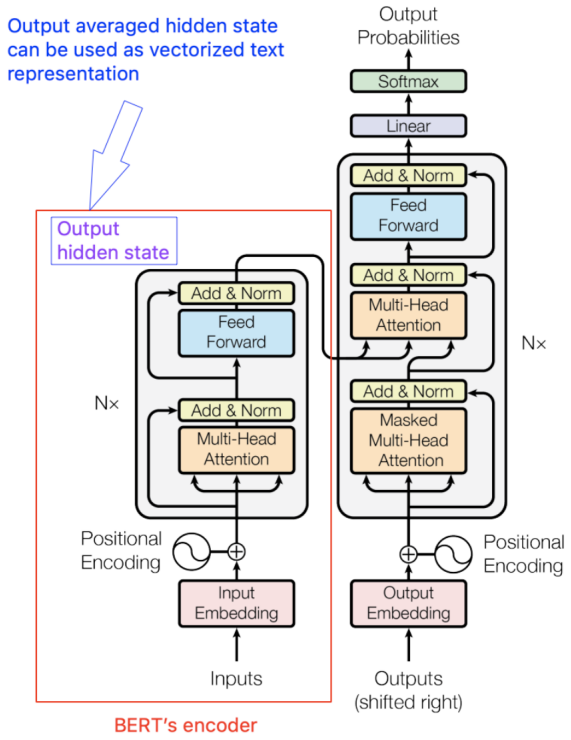

Figure 8. *BERT working scheme [8]*

The Transformer architecture excels at handling text data that is inherently sequential. They take a text sequence as input and produce another as output, such as translating an input English sentence to Spanish.

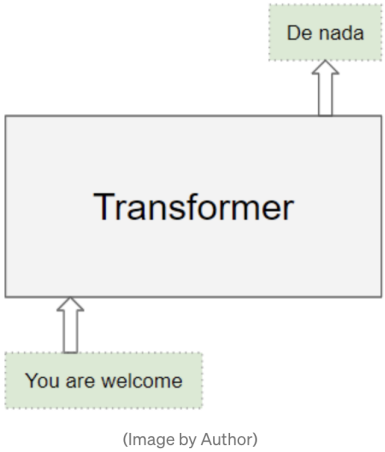

Figure 9. *Illustration of transformer architecture*

At its core, it contains a stack of Encoder and Decoder layers. To avoid confusion, we will refer to the individual layer as

an Encoder or a Decoder and use an Encoder or Decoder stack for a group of Encoder layers. The Encoder and Decoder stacks have corresponding Embedding layers for their respective inputs. Finally, there is an Output layer to generate the final output.

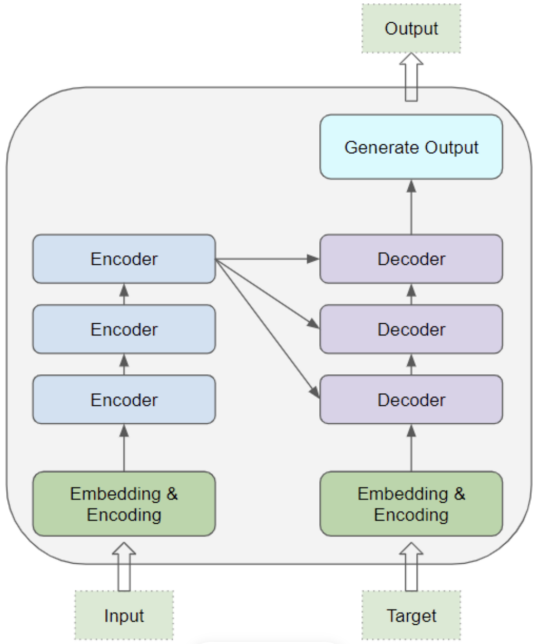

Figure 10. *Illustration of Encoder and Decoder stack layers*

All the Encoders are identical to one another. Similarly, all the Decoders are identical. The encoder contains the all-important Self-attention layer that computes the relationship between different words in the sequence and a Feed-forward layer.

The encoder contains the all-important Self-attention layer that computes the relationship between different words in the sequence and a Feed-forward layer. The decoder contains the Self-attention layer, the Feed-forward layer, and a second Encoder-Decoder attention layer. Each Encoder and Decoder has its own set of weights. The encoder is a reusable module that is the defining component of all Transformer architectures. In addition to the above two layers, it also has Residual skip connections around both layers and two LayerNorm layers. There are many variations of the Transformer architecture. For example, some Transformer architectures have no decoder and rely only on the encoder. The Transformer works slightly differently during Training and while making Inferences. Let us first look at the flow of data during Training. Training data consists of two parts: The source or input sequence (e.g. "You are welcome" in English for a translation problem) The destination or target sequence (e.g. "De nada" in Spanish) The Transformer aims to learn how to output the target sequence using both the input and target sequence.

How text-to-vector conversion works

BERT consists of two parts: encoder and decoder. It was trained on masked language modeling and next-sentence prediction tasks. The encoder has hidden output states. We can say that averaging these vectors represents a vector representation of a whole text. So this representation can be used as a feature vector of this text.

Reddit practical use case explanation

The main goal of the use case is to calculate a similarity score between two Reddit users by analyzing their posts and comments. The similarity score of different pairs of users can be compared and reflects their similarity. The project was created in Jupyter notebook on Python using the following modules:

1. PRAW is a Python Reddit API wrapper that helps create bots and scripts for Reddit.
2. Itertools is used for iterating over data structures that can be bypassed with a for-loop.
3. Torch is a machine learning open-source library.
4. Transformers holds pre-trained models for NLP (Natural Language Processing)

```
from transformers import
BertTokenizerFast, BertTokenizer,
BertModel
import praw
from itertools import chain
import torch
```

A centralized logging system was enabled in Transformers, and a setup for verbosity errors was also added.

```
from transformers import logging
logging.set_verbosity_error()
```

Application classes

1. RedditUser: class for fetching user texts (comments and posts)

```
class RedditUser:
    def __init__(self, reddit: praw.Reddit,
    user_name):
        self.user = reddit.redditor(user_name)
    def fetch_texts(self, depth):
        return
        chain(self._fetch_comments(depth),
                    self._fetch_post(depth))
    def _fetch_comments(self, depth):
        return map(lambda c: c.body,
    self.user.comments.new(limit=depth))
    def _fetch_post(self, depth):
        return map(lambda s: s.selftext,
    self.user.submissions.new(limit=depth))
```

2. RedditAuth and Reddit: classes for working with Reddit API. Delegates fetching to Reddituser

```
class RedditAuth:
  def __init__(self, client_id, client_secret,
  user_agent):
    self.client_id = client_id
    self.client_secret = client_secret
    self.user_agent = user_agent
class Reddit:
  def __init__(self, text_fetch_depth,
  reddit_auth):
    self.text_fetch_depth = text_fetch_depth
    self.reddit = praw.Reddit(
      client_id=reddit_auth.client_id,
      client_secret=reddit_auth.client_secret,
```

```
        user_agent=reddit_auth.user_agent,
        check_for_async=False)
    def user_texts(self, user_name):
      return RedditUser(self.reddit,
      user_name).fetch_texts
      (self.text_fetch_depth)
```

3. Classes extToVecUsingHiddenState and TextToVecUsing-PoolerOutput: for 2 different ways for converting a text to a vector
4. Application: a class for starting the project. It also contains some utility methods.

Project input parameters

1. A Reddit app credentials: client_id, client_secret, user_agent.
2. Parameter text_fetch_depth describes how many new comments and posts will be fetched. This parameter influences RAM and time consumption.
3. Parameter max_tokenizer_length is responsible for creating a token for sequences of the model. This parameter influences RAM and time consumption.

How does the application work?

The system asked users to enter two different usernames for whom we could calculate a similarity score. The Similarity Score can be defined as a probability between 0 and 100 percent, where 0 is no matches of the submitted text, and 100 is that the text is fully similar. For each user, we gain information from the open Reddit API:

1. Downloads last: text_fetch_depth comments and posts of the users using the PRAW Python library,
2. Transforms these texts to vectors applying BERT Tokenizer and BERT pre-trained encoder,
3. Averaging user text vectors to build a vector characterizing this user.

Ultimately, we calculate the scalar product of 2 vectors using the cosine similarity score.

```
class Application:
  def __init__(self,
               text_to_vec,
               text_fetch_depth,
               reddit_auth):
      self.reddit = Reddit(text_fetch_depth,
      reddit_auth)
      self.text_to_vec = text_to_vec

  def _mean(self, list_of_text_vec):
      matrix_of_text_vectors
      = torch.stack(list_of_text_vec, dim=0)
  return torch.mean(matrix_of_text_vectors,
  dim=0)

  def _calculate_similarity_score(self,
  user_name_1,   user_name_2):
    list_of_text_vec_1 = list(map(self.text_to_vec,
    self.reddit.user_texts(user_name_1)))
    vec_1 = self._mean(list_of_text_vec_1)
    list_of_text_vec_2 = list(map(self.text_to_vec,
    self.reddit.user_texts(user_name_2)))
    vec_2 = self._mean(list_of_text_vec_2)
    return torch.dot(vec_1, vec_2) / (vec_1.norm()
    * vec_2.norm())

  @staticmethod
  def main_loop(text_to_vec, text_fetch_depth,
  reddit_auth):
    app = Application(text_to_vec,
    text_fetch_depth, reddit_auth)

while True:
user_name_1 = input('Please enter first
user name: ')
user_name_2 = input('Please enter second
user name: ')
print(f'''The similarity score
between {user_name_1} and {user_name_2}
is {app._calculate_similarity_score(user_name_1,
user_name_2)}''')
The outcomes of the model of the system look as
a similarity score is in the range  [0.0, 1.0]:
Please enter the first user name: xtilexx
Please enter the second user name: Sir_Loinbeef
The similarity score between xtilexx and
Sir_Loinbeef
            is 0.9693130254745483
Please enter the first user name: Repulsive_Love_
Please enter the second user name: ForecastForFourCats
The similarity score between Repulsive_Love_ and
ForecastForFourCats
            is 0.6157388091087341
Please enter the first user name: xtilexx
Please enter the second user name: Repulsive_Love_
The similarity score between xtilexx
and Repulsive_Love_
            is 0.8943862915039062
Please enter the first user name: xtilexx
Please enter the second user name: ForecastForFourCats
The similarity score between xtilexx and
ForecastForFourCats
            is 0.8847013115882874
```

Project workflow

For each user:

1. Downloads last text_fetch_depth comments and posts of the users using the PRAW python library.
2. Transforms these texts to vectors using BERT Tokenizer and BERT pre-trained encoder.
3. Averaging user text vectors to build a vector characterizing this user.

At the end of our vectors, we calculate the cosine similarity score. Cosine similarity [10] is a measure of similarity between two sequences of numbers. To define it, the sequences are viewed as vectors in an inner product space, and the cosine similarity is defined as the cosine of the angle between them, that is, the dot

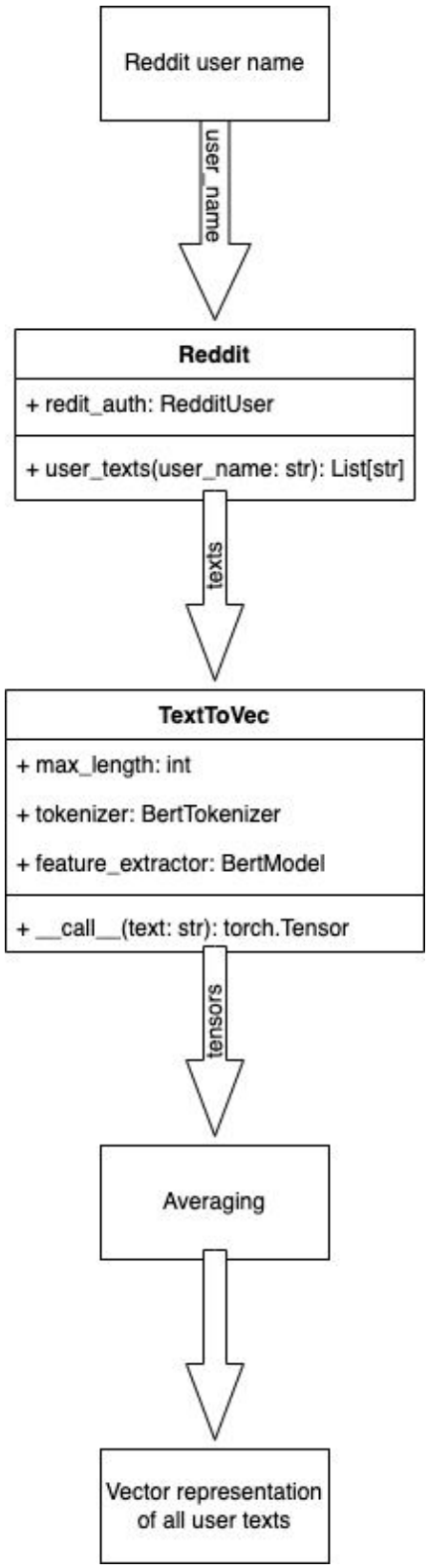

Figure 11. *Building user vector representation*

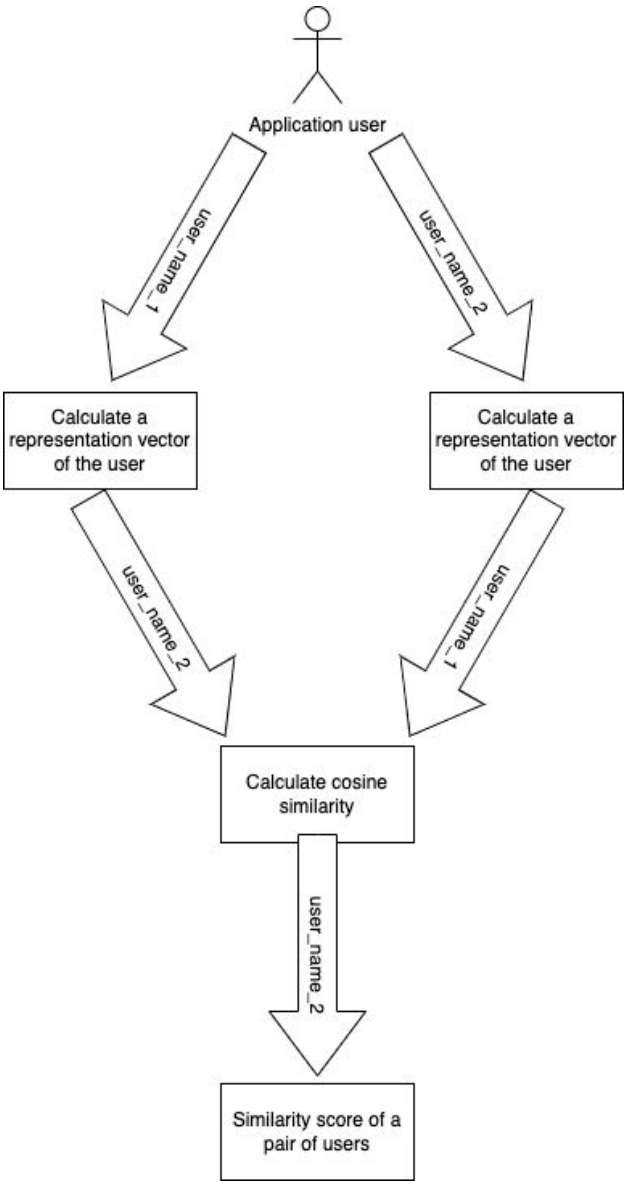

Figure 12. *Overall workflow*

product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the magnitudes of the vectors but only on their angle. The cosine similarity always belongs to the interval. For example, two proportional vectors have a cosine similarity of 1, two orthogonal vectors have a similarity of 0, and two opposite vectors have a similarity of -1. The cosine similarity is used in positive space, where the outcome is neatly bounded. In information retrieval and text mining, each word is assigned a different coordinate, and the vector of the number of occurrences of each word in the document represents a document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter and independently of the length of the documents. The technique is also used to measure cohesion within clusters in data mining. One advantage of cosine similarity is its low complexity, especially for sparse vectors: only the non-zero coordinates

need to be considered. The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$A * B = \mid A \mid\mid B \mid \cos\theta. (1)$$

Given two vectors of attributes, A and B, the cosine similarity, $\cos(\theta)$, is represented using a dot product and magnitude as:

$$\text{cosine similarity} = S_c(A,B) := \cos\theta = \frac{A * B}{\mid A \mid\mid B \mid} =$$

$$= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, (2)$$

where A_i and B_i are components of a vector.

Future work and summary

This paper presents a practical Open Source Intelligence (OSINT) use case for user similarity measurements using open profile data from the Reddit social network. The approach was invented using a pre-trained BERT model.

We can extend our model and calculate feature vectors for subreddits. In that way, we can find similar to the user's subreddits and recommend them to him.

The results of the model are as the following:

- 1. Using the last hidden state. It performs with good results. Here we used an averaging of all hidden states of the BERT encoder: sequence output, last layer of the model. Each hidden state represents an internal representation of a token in an input sequence of words. The similarity score is in the range [0.5, 1.0].
- 2. Using pooler output. Additional 'pooler' layer, aggregation of all hidden states of all input tokens: BertPooler. That approach needs more time for investigation.

Could we have tried the well-known algorithm Word2vec as a word encoder? We could map each word in a text to a vector using Word2vec. Then all these vectors could be aggregated by averaging. Unfortunately, this method contains an obvious flaw: It evaluates words individually rather than considering a text as a sequence of words with inner structural meaning.

As a proposal for suggestions for improvements and future work could be:

- Improving preprocessing stage of Reddit posts and comments (delete links, emojis, etc.)
- Fine-tuning of hyperparameters of BERT model for project needs

Acknowledgment

This work was supported partially by the European Union in the framework of ERASMUS MUNDUS, Project CyberMACS (Project #101082683) (`https://cybermacs.eu`).

References

[1] Hassan NA, Hijazi R "Open source intelligence methods and tools" – A practical guide to Online intelligence. APress, Berkeley, 2018

[2] Pais, V. F., & Ciobanu, D. S. OSINT for B2B platforms. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2014

[3] Jiafu Tang, Jiwei Song, Discrete particle swarm optimization combined with no-wait algorithm in stages for scheduling mill roller annealing process, 2010

[4] YouTube. (Nov 28, 2017). Youtube video: Attention Is All You Need by Yannic Kilcher - `https://www.youtube.com/watch?v=iDulhoQ2pro&ab_channel=YannicKilcher`,(Last access: December 25, 2022)

[5] Transformers Explained Visually (Part 1): Overview of Functionality by Ketan Doshi

[6] Brian Dean: Reddit User and Growth Stats (Updated Oct 2021) `https://backlinko.com/reddit-users`, (Last access: December 25, 2022)

[7] OSINT Framework: The Perfect Cybersecurity Intel Gathering Tool `https://securitytrails.com/blog/osint-framework`, (Last access: December 25, 2022)

[8] Adrian Tam: A Brief Introduction to BERT. `https://machinelearningmastery.com/a-brief-introduction-to-bert/`, (Last access: December 25, 2022)

[9] BERT Transformers – How Do They Work? March 29, 2021 `https://www.exxactcorp.com/blog/Deep-Learning/how-do-bert-transformers-work`, (Last access: December 25, 2022)

[10] Wikipedia: Cosine similarity. `https://en.wikipedia.org/wiki/Cosine_similarity`, (Last access: December 25, 2022)

[11] OSINT Framework. `https://osintframework.com/`, (Last access: December 25, 2022)

[12] AlertMedia, "What Businesses Need to Know About Open Source Intelligence (OSINT)" `https://www.alertmedia.com/blog/open-source-intelligence/` (Last access: Nov. 12, 2022).

[13] Signal, "The Pivotal Role of OSINT for Effective Emergency Management Pivotal." `https://www.getsignal.info/blog/osint-emergency-management` (Last access: Nov. 12, 2022).

[14] Qadir, Junaid & Ali, Anwaar & Rasool, Raihan & Zwitter, Andrej & Sathiaseelan, Arjuna & Crowcroft, Jon. (2016): "Crisis Analytics: Big Data Driven Crisis Response." Journal of International Humanitarian Action. 1. `https://doi.org/10.1186/s41018-016-0013-9`, (Last access: Nov. 12, 2022).

[15] Echosec SYSTEMS LTD, "6 Reasons Why Open-Source Intelligence is Climbing the Priority Ladder" `https://www.echosec.net/blog/6-reasons-why-open-source-intelligence-is-climbing-the-priority-ladder` (Last access: Nov. 13, 2022).

[16] TechTarget, "Crisis Management" `https://www.techtarget.com/whatis/definition/crisis-management` (Last access: Nov. 15, 2022).

[17] InaSAFE, `http://inasafe.org/home/index.html` (Last access: Nov. 17, 2022).

[18] Australian Government, "New disaster management software released worldwide," `https://www.ga.gov.au/news-events/news/latest-news/new-disaster-`

management-software-released-worldwide (Last access: Nov. 17, 2022).
[19] Pranantyo, Ignatius & Fadmastuti, Mahardika. (2014). "InaSAFE applications in disaster preparedness." https://doi.org/10.1063/1.4915053.
[20] JSTOR, "4 Case Study 3 – Ushahidi, An Open Platform for Situation Awareness" https://www.jstor.org/stable/resrep12574.7?seq=1 (Last access: Nov. 17, 2022).
[21] Ushahidi, "Crisis Mapping Haiti: Some Final Reflections" https://www.ushahidi.com/about/blog/crisis-mapping-haiti-some-final-reflections (Last access: Nov. 17, 2022).
[22] Wald, D.J., Worden, C.B., Thompson, E.M., Hearne, M. (2019). "Earthquakes, ShakeMap." In: Gupta, H. (eds) Encyclopedia of Solid Earth Geophysics. Encyclopedia of Earth Sciences Series. Springer, Cham. https://doi.org/10.1007/978-3-030-10475-7_182-1, (Last access: Nov. 17, 2022).
[23] reliefweb, "USGS ShakeMap: Haiti Region - Tue Jan 12, 2010 21:53:09 GMT" https://reliefweb.int/map/haiti/usgs-shakemap-haiti-region-tue-jan-12-2010-215309-gmt (Last access: Nov. 17, 2022).
[24] Verified Market Research, "Open Source Intelligence (OSINT) Market Size And Forecast" https://www.verifiedmarketresearch.com/product/open-source-intelligence-osint-market/ (Last access: Nov. 19, 2022).
[25] USGS,"ShakeMap" https://earthquake.usgs.gov/data/shakemap/ (Last access: Nov. 19, 2022)
[26] Github, https://github.com/usgs/shakemap (Last access: Nov. 19, 2022)
[27] Ida Norheim-Hagtun, Patrick Meier: "Crowdsourcing for Crisis Mapping in Haiti." Innovations: Technology, Governance, Globalization 2010; 5 (4): 81–89. https://doi.org/10.1162/INOV_a_00046
[28] Dave Chaffey, "Global social media statistics research summary 2022" https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (Last access: Nov. 19, 2022)
[29] Sahana Foundation, "Eden" https://sahanafoundation.org/products/eden/ (Last access: Nov. 20, 2022)
[30] Talkwalker, "Try the best crisis management tools" https://www.talkwalker.com/blog/best-crisis-management-tools (Last access: Nov. 20, 2022)
[31] CSIRO, "ESA: Software for Emergency Situation Awareness" https://data61.csiro.au/en/Our-Research/Our-Work/Safety-and-Security/Disaster-Management/ESA (Last access: Nov. 18, 2022)
[32] CSIRO, "Emergency Situation Awareness" https://www.csiro.au/en/research/technology-space/ai/emergency-situation-awareness (Last access: Nov. 19, 2022)
[33] CSIRO ESA, "AUS tweets analysed by ESA over last 4 days" https://esa.csiro.au/aus/index.html (Last access: Nov. 17, 2022)
[34] Statista, "Number of deaths from natural disaster events globally from 2007 to 2021" https://www.statista.com/statistics/510952/number-of-deaths-from-natural-disasters-globally/ (Last access: Nov. 20, 2022)
[35] Our World in Data, "Global reported natural disasters by type, 1970 to 2019" https://ourworldindata.org/grapher/natural-disasters-by-type (Last access: Nov. 20, 2022).
[36] Our World in Data, "Natural Disasters by Hannah Ritchie and Max Roser" https://ourworldindata.org/natural-disasters#:~:text=Natural%20disasters%20kill%20on%20average,from%200.01%25%20to%200.4%25 (Last access: Nov. 20, 2022)
[37] Australian Government, "Australian Government Crisis Management Framework (AGCMF)" https://www.pmc.gov.au/resource-centre/national-security/australian-government-crisis-management-framework, (Last access: Nov. 21, 2022)
[38] Australian Government, "The use of social media in countrywide disaster risk reduction public awareness strategies" https://knowledge.aidr.org.au/resources/ajem-jan-2015-the-use-of-social-media-in-countrywide-disaster-risk-reduction-public-awareness-strategies/ (Last access: Nov. 20, 2022)
[39] Cabinet Office, "Disaster Management in Japan" https://www.bousai.go.jp/en/documentation/white_paper/pdf/2021/R3_hakusho_english.pdf (Last access: Nov. 21, 2022).
[40] National Library of Medicine, "Disaster Management in Japan" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059167/ (Last access: Nov. 21, 2022).
[41] PEARY, Brett & Shaw, Rajib & TAKEUCHI, Yukiko. (2012). "Utilization of Social Media in the East Japan Earthquake and Tsunami and its Effectiveness." Journal of Natural Disaster Science. 34. 3-18. https://doi.org/10.1186/s41018-016-0013-9, (Last access: Nov. 21, 2022).
[42] B. Birregah et al., "Multi-layer Crisis Mapping: A Social Media-Based Approach," 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2012, pp. 379-384, https://ieeexplore.ieee.org/document/6269762, (Last access: Nov. 20, 2022).
[43] ACM Digital Library, "A sensitive Twitter earthquake detector" https://dl.acm.org/doi/10.1145/2487788.2488101, (Last access: Nov. 20, 2022)
[44] Schwarz, Klaus; Franziska Schwarz, Reiner Creutzburg: "Conception and implementation of professional laboratory exercises in the field of open source intelligence (OSINT)". Proceed. Electronic Imaging Symposium 2020 (San Francisco, USA), Mobile Devices and Multimedia: Technologies, Algorithms & Applications Conference (MOBMU) 2020, https://doi.org/10.2352/ISSN.2470-1173.2020.3.MOBMU-278, (Last access: Nov. 22, 2022).
[45] Schwarz, Klaus; Reiner Creutzburg: "Design of Professional Laboratory Exercises for Effective State-of-the-Art OSINT Investigation Tools - Part 1: RiskIQ PassiveTotal". Proceed. Electronic Imaging Symposium 2021 (San Francisco, USA), Mobile Devices and Multime-

dia: Technologies, Algorithms & Applications Conference (MOBMU) 2021, https://doi.org/10.2352/ISSN.2470-1173.2021.3.MOBMU-043, Last access: Nov. 22, 2022).

[46] Schwarz, Klaus; Reiner Creutzburg: "Design of Professional Laboratory Exercises for Effective State-of-the-Art OSINT Investigation Tools - Part 2: Censys". Proceed. Electronic Imaging Symposium 2021 (San Francisco, USA), Mobile Devices and Multimedia: Technologies, Algorithms & Applications Conference (MOBMU) 2021, https://doi.org/10.2352/ISSN.2470-1173.2021.3.MOBMU-044, Last access: Nov. 22, 2022.

[47] Schwarz, Klaus; Reiner Creutzburg: "Design of Professional Laboratory Exercises for Effective State-of-the-Art OSINT Investigation Tools - Part 3: Maltego". Proceed. Electronic Imaging Symposium 2021 (San Francisco, USA), Mobile Devices and Multimedia: Technologies, Algorithms & Applications Conference (MOBMU) 2021, https://doi.org/10.2352/ISSN.2470-1173.2021.3.MOBMU-045, Last access: Nov. 22, 2022.

[48] Kant, Daniel; Reiner Creutzburg: 'Investigation of risks for Critical Infrastructures due to the exposure of SCADA systems and industrial controls on the Internet based on the search engine Shodan". Proceed. Electronic Imaging Symposium 2020 (San Francisco, USA), Mobile Devices and Multimedia: Technologies, Algorithms & Applications Conference (MOBMU) 2020, https://doi.org/10.2352/ISSN.2470-1173.2020.3.MOBMU-253, Last access: Nov. 22, 2022.

[49] M. S. Wong, N. Hideki and N. Yasuyuki, "The Incorporation of Social Media in an Emergency Supply and Demand Framework in Disaster Response," 2018 IEEE Intl. Conf. on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications https://ieeexplore.ieee.org/document/8672243, (Last access: Nov. 20, 2022).

[50] T. Sakaki et al., "The possibility of social media analysis for disaster management," 2013 IEEE Region 10 Humanitarian Technology Conference, 2013, pp. 238-243, https://www.scopus.com/record/display.uri?eid=2-s2.0-84893406250&origin=inward&txGid=7adf7d88a2a5fe170927ab1110f2009f, (Last access: Nov. 20, 2022).

Author Biography

Valeriya Vishnevskaya is currently pursuing a master's degree in Artificial Intelligence and Big Data at the SRH Hochschule Berlin. She worked as an AI intern at the motorsport division of Bosch Engineering GmbH, Germany. She is currently writing a master's thesis with Bosch. She received her MTech degree in Informatics and Computer Engineering from Moscow Aviation University in 2016.

Klaus Schwarz received his B. Sc. and M.Sc. in Computer Science from Brandenburg University of Applied Sciences (Germany) in 2017 and 2020, respectively. He is currently a Ph.D. student at the University of Granada, Spain. His research interests include IoT and smart home security, OSINT, mechatronics, additive manufacturing, embedded systems, artificial intelligence, and cloud security. As a faculty member, he is developing a graduate program in Applied Mechatronic Systems focusing on Embedded Systems at SRH Berlin University of Applied Sciences.

Reiner Creutzburg is a Retired Professor for Applied Computer Science at the Technische Hochschule Brandenburg in Brandenburg, Germany. Since 2019 he has been a Professor of IT Security at the SRH Berlin University of Applied Sciences, Berlin School of Technology. He is a member of the IEEE and SPIE and chairman of the Multimedia on Mobile Device (MOBMU) Conference at the Electronic Imaging conferences since 2005. In 2019, he was elected a member of the Leibniz Society of Sciences to Berlin e.V. His research interest is focused on Cybersecurity, Digital Forensics, Open Source Intelligence (OSINT), Multimedia Signal Processing, eLearning, Parallel Memory Architectures, and Modern Digital Media and Imaging Applications.