Report | Project Part 3

The project has been published on GitHub at the following URL and is available within this specific tag:

URL: https://github.com/ericdalmases/IRWA.git
Tag: IRWA-2023-part-3
(https://github.com/ericdalmases/IRWA/releases/tag/IRWA-2023-part-3)

Since this deliverable is a continuation of the previous one, some of the code of the previous deliverables has been added in this delivery.

All the tweets that are shown as output have been checked to contain all the terms of the query, to fulfill the AND requirement of all the queries of this lab.

Our own score

To generate our own scoring system, we initially developed a new class named "User" to house all the relevant variables extracted from each tweet's user. These variables were crucial in the formulation of the new score. We then assessed the variables to identify those with potential impact on the ranking. Upon careful consideration, we determined the order of significance for the variables associated with each tweet as follows: the tweet's retweet count, the number of likes it received, the count of replies, the follower-to-followed ratio (with preference given to accounts with significantly fewer followed compared to followers), the account's verification status, and the number of tweets (with a negative valuation for accounts with excessively high tweet counts). After adding all the weights the scoring equation would look like this:

```
2*num_retweets + num_likes + 1.5*num_replies +
log(followers+1)/log(max(10,num_followed)) + verified + (1/num_tweets)
```

We also considered taking into account whether a tweet was a reply tweet but since we only had 82 of them we decided to not do so.

To resolve the problem of disproportionately high values for specific users, we introduced a logarithmic ratio. This adjustment ensures that other attributes are not overshadowed and maintain a meaningful impact in the score calculation. Then we also needed to make sure that the denominator of the fraction was at least 1 and that no calculation could be invalid.

The current ranking system significantly differs from the previous one, which relied on tf-idf. Previously, the evaluation was solely based on the text content of each tweet, considering factors such as word count and the frequency of query terms. In contrast, the updated approach considers the broader influence of a tweet and its user within the social network. Factors such as the number of likes a tweet receives and the overall influence of the account are now taken into account, providing a more comprehensive assessment of the tweet's relevance. Each method has its pros and cons:

tf-idf based score:

Pros:

- **Textual Relevance**: tf-idf is effective at capturing the textual relevance of a document by considering the frequency of specific terms. This makes it suitable for tasks where the focus is on the content itself.
- **Simplicity**: The tf-idf algorithm is relatively simple to implement and computationally efficient. It's a straightforward approach for text-based tasks.

Cons:

- **Limited Context**: tf-idf may struggle with capturing the context and semantics of the text. It treats each term independently, potentially missing the nuances of language and meaning. It also does not take into account the possible relevance of the tweet: a more influential tweet might be more relevant than another one in this subject, even if its tf-idf is lower.

Our score:

Pros

- **Contextual Relevance**: This method captures the broader context by considering social network impact, user influence, and engagement metrics. It provides a more holistic view of a tweet's significance.
- **Adaptability**: The social network impact method is adaptable to changes in user behavior and evolving trends on social platforms. It can better reflect real-time dynamics.

Cons

- **Data Availability Dependency**: Success with this method relies on the availability and accuracy of social network metrics. In cases where such data is sparse or unreliable, the performance of the algorithm may be compromised.
- **Complexity**: Integrating social network impact introduces complexity into the ranking system. Managing and interpreting diverse metrics such as likes, retweets, and user influence requires careful consideration and may add computational overhead.

Below, in Figure 1, 2, 3, 4 and 5, there are the top 20 rankings obtained using TF-IDF + cosine similarity and our-score + cosine similarity for each query. The queries are centered around topics such as *conflict in Ukraine* (Figure 1), *gas in Ukraine* (Figure 2), *president of Russia* (Figure 3), *support in Ukraine* (Figure 4), and *power of Russia* (Figure 5). We have decided to just print the doc_id for readability. More information about the tweets can be found in the code.

Analyzing Figure 1, we observe that **fifteen** tweets appear in both rankings, signifying a noteworthy overlap. However, it's notable that the position of these tweets in the rankings doesn't align.

For the second query, the number of tweets that are in both rankings decreases to **eleven**. The number of coincidences are not as significant as the first query, but taking into account that to compute the cosine similarity we only used our score, not the tf-idf score it is still quite surprising. As in query 1, and it will be extrapolated to all the queries, the position in both rankings will always differ.

For the third query, we have again **eleven** tweets which coincide in both rankings.

Moving to Figure 4, the query with the least number of overlapping tweets is encountered

In the final query (*power of Russia*), Figure 5 stands out with the highest number of overlapping tweets—**eighteen** in total. This result is particularly impressive since our score neglects word frequency and tweet length, factors integral to TF-IDF. The substantial matching between rankings is surprising in this context.

```
Top 20 results out of 2659 for the searched query:
page_id= doc_3754
page_id= doc_2444
page_id= doc_698
page_id= doc_689
page_id= doc_1078
page_id= doc_1077
page_id= doc_324
page_id= doc_3057
page_id= doc_3645
page_id= doc_2852
page_id= doc_1505
page_id= doc_2469
page_id= doc_3756
page_id= doc_2132
page_id= doc_1105
page_id= doc_2798
page_id= doc_1709
page_id= doc_2766
page_id= doc_2251
page_id= doc_1764
```

```
Top 20 results out of 2659 for the searched query:
page_id= doc_1709
page_id= doc_3645
page_id= doc_2852
page_id= doc_1077
page_id= doc_2766
page_id= doc_3759
page_id= doc_324
page_id= doc_2798
page_id= doc_1105
page_id= doc_1029
page_id= doc_689
page_id= doc_2049
page_id= doc_3905
page_id= doc_1505
page_id= doc_445
page_id= doc_2251
page_id= doc_2469
page_id= doc_2132
page_id= doc_3756
page_id= doc_1764
```

Figure 1 (tf-idf left and our-score right)

```
Top 20 results out of 2674 for the searched query:
page_id= doc_220
page_id= doc_3209
page_id= doc_3080
page_id= doc_3636
page_id= doc_2828
page_id= doc_2895
page_id= doc_2103
page_id= doc_3011
page_id= doc_1391
page_id= doc_3747
page_id= doc_3692
page_id= doc_3002
page_id= doc_3772
page_id= doc_3923
page_id= doc_2681
page_id= doc_3913
page_id= doc_3441
page_id= doc_3926
page_id= doc_2862
page_id= doc_2484
```

```
Top 20 results out of 2674 for the searched query:
page_id= doc_2886
page_id= doc_1333
page_id= doc_3036
page_id= doc_423
page_id= doc_1075
page_id= doc_220
page_id= doc_2210
page_id= doc_2828
page_id= doc_850
page_id= doc_3441
page_id= doc_3926
page_id= doc_2484
page_id= doc_1452
page_id= doc_2492
page_id= doc_2862
page_id= doc_3080
page_id= doc_3692
page_id= doc_3923
page_id= doc_3913
page_id= doc_2681
```

Figure 2 (tf-idf left and our-score right)

```
Top 20 results out of 1662 for the searched query:
page_id= doc_3996
page_id= doc_408
page_id= doc_632
page_id= doc_403
page_id= doc_3473
page_id= doc_2393
page_id= doc_453
page_id= doc_417
page_id= doc_3692
page_id= doc_1893
page_id= doc_507
page_id= doc_1241
page_id= doc_585
page_id= doc_228
page_id= doc_470
page_id= doc_87
page_id= doc_1967
page_id= doc_418
page_id= doc_30
page_id= doc_1771
```

```
Top 20 results out of 1662 for the searched query:
page_id= doc_418
page_id= doc_632
page_id= doc_403
page_id= doc_470
page_id= doc_1197
page_id= doc_87
page_id= doc_3996
page_id= doc_233
page_id= doc_3078
page_id= doc_179
page_id= doc_417
page_id= doc_1196
page_id= doc_156
page_id= doc_2938
page_id= doc_962
page_id= doc_585
page_id= doc_247
page_id= doc_453
page_id= doc_408
page_id= doc_2393
```

Figure 3 (tf-idf left and our-score right)

```
Top 20 results out of 2674 for the searched query:
page_id= doc_3532
page_id= doc_3020
page_id= doc_3277
page_id= doc_3094
page_id= doc_939
page_id= doc_844
page_id= doc_324
page_id= doc_1298
page_id= doc_376
page_id= doc_1416
page_id= doc_3120
page_id= doc_3054
page_id= doc_895
page_id= doc_696
page_id= doc_2817
page_id= doc_3978
page_id= doc_1780
page_id= doc_1442
page_id= doc_627
page_id= doc_3740
```

```
Top 20 results out of 2674 for the searched query:
page_id= doc_2715
page_id= doc_3094
page_id= doc_3152
page_id= doc_1421
page_id= doc_1442
page_id= doc_2817
page_id= doc_124
page_id= doc_1416
page_id= doc_111
page_id= doc_3699
page_id= doc_3725
page_id= doc_3110
page_id= doc_3858
page_id= doc_324
page_id= doc_3228
page_id= doc_1298
page_id= doc_627
page_id= doc_330
page_id= doc_150
page_id= doc_1839
```

Figure 4 (tf-idf left and our-score right)

```
Top 20 results out of 1639 for the searched query:
page_id= doc_2554
page_id= doc_3090
page_id= doc_2705
page_id= doc_1254
page_id= doc_1253
page_id= doc_1251
page_id= doc_1117
page_id= doc_2548
page_id= doc_2478
page_id= doc_2316
page_id= doc_1446
page_id= doc_3660
page_id= doc_457
page_id= doc_3760
page_id= doc_1741
page_id= doc_741
page_id= doc_633
page_id= doc_1260
page_id= doc_2611
page_id= doc_3572
```

```
Top 20 results out of 1639 for the searched query:
page_id= doc_633
page_id= doc_2554
page_id= doc_2611
page_id= doc_741
page_id= doc_2316
page_id= doc_1446
page_id= doc_3760
page_id= doc_2548
page_id= doc_2478
page_id= doc_1117
page_id= doc_1253
page_id= doc_1260
page_id= doc_3090
page_id= doc_3572
page_id= doc_2602
page_id= doc_1744
page_id= doc_1251
page_id= doc_1254
page_id= doc_457
page_id= doc_1741
```

Figure 5 (tf-idf left and our-score right)

Word2Vec & Cosine Similarity

First of all we trained our own Word2Vec model with all the tweet tokens we had, as in the previous deliverable. We decided to keep a vector size of 100 for each token and a window size of 5. The sequences of tokens we used to perform the training were the tokenized tweets of our database.

Once we had the model trained we started a very simple iterative process. We embedded each query and all the tweets that contained all the query terms. Then, we computed the cosine similarity between the query and the filtered tweets. So, we had 5 lists (one per query) of tweets which were sorted by similarity between them and the query. Finally, we selected the top 20 tweets for each of the queries.

The results we obtained are the following:

For *conflict in Ukraine*: *gas in Ukraine* (Figure 2), *president of Russia* (Figure 3), *support in Ukraine* (Figure 4), and *power of Russia* (Figure 5).

page_id= doc_2251 – page_title: https://www.twitter.com/W_W_3_2022/status/1575525609714831373
page_id= doc_2469 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575473073507155969
page_id= doc_1505 – page_title: https://www.twitter.com/Shadi_Alkasim/status/1575692522247987201
page_id= doc_3057 – page_title: https://www.twitter.com/YugoUnderground/status/1575329009377869824
page_id= doc_3754 – page_title: https://www.twitter.com/KacoBlokland/status/1575182921664671746
page_id= doc_689 – page_title: https://www.twitter.com/_Policy_Center/status/1575825276147490816
page_id= doc_698 – page_title: https://www.twitter.com/_Policy_Center/status/1575825011776290817
page_id= doc_324 – page_title: https://www.twitter.com/_Nex3_/status/1575879420300390403
page_id= doc_2852 – page_title: https://www.twitter.com/ttindia/status/1575380893132201984
page_id= doc_1078 – page_title: https://www.twitter.com/TechUnityInc/status/1575796243036803072
page_id= doc_1077 – page_title: https://www.twitter.com/QAValley/status/1575796260639977472
page_id= doc_2049 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575582530899562496
page_id= doc_3756 – page_title: https://www.twitter.com/Suspended_Acct/status/1575182900886384640
page_id= doc_3765 – page_title: https://www.twitter.com/khalilxxx990/status/1575181650178473985
page_id= doc_3759 – page_title: https://www.twitter.com/khalilxxx990/status/1575182363268911104
page_id= doc_1105 – page_title: https://www.twitter.com/EulogyForEurope/status/1575793337222586368
page_id= doc_2444 – page_title: https://www.twitter.com/MatiStein/status/1575476963833434113
page_id= doc_3905 – page_title: https://www.twitter.com/Suspended_Acct/status/1575164722244358144
page_id= doc_3645 – page_title: https://www.twitter.com/PatilSushmit/status/1575194020690927616
page_id= doc_2798 – page_title: https://www.twitter.com/frontline_india/status/1575402501813276675

Figure 6

page_id= doc_3636 – page_title: https://www.twitter.com/musielak_/status/1575194718807289856
page_id= doc_3923 – page_title: https://www.twitter.com/W_W_3_2022/status/1575162638174081024
page_id= doc_3080 – page_title: https://www.twitter.com/Mickey17176/status/1575318568463466497
page_id= doc_220 – page_title: https://www.twitter.com/marydejevsky/status/1575895067868405764
page_id= doc_2828 – page_title: https://www.twitter.com/toddxz/status/1575390531567308801
page_id= doc_2681 – page_title: https://www.twitter.com/HAL96661/status/1575440186581962752
page_id= doc_850 – page_title: https://www.twitter.com/publicistjourn/status/1575816049743699968
page_id= doc_3378 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575239459481784328
page_id= doc_2895 – page_title: https://www.twitter.com/KHonkonen/status/1575372481136918528
page_id= doc_3441 – page_title: https://www.twitter.com/balog_amy/status/1575228294085365760
page_id= doc_3926 – page_title: https://www.twitter.com/W_W_3_2022/status/1575162188003627015
page_id= doc_3747 – page_title: https://www.twitter.com/sustellers/status/1575183567595839488
page_id= doc_2484 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575470394244571136
page_id= doc_2210 – page_title: https://www.twitter.com/Chronology22/status/1575535400864677896
page_id= doc_3913 – page_title: https://www.twitter.com/W_W_3_2022/status/1575164055588700177
page_id= doc_1075 – page_title: https://www.twitter.com/ActForUA/status/1575796555172306944
page_id= doc_1452 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575709956699045888
page_id= doc_3002 – page_title: https://www.twitter.com/VoskopoulosPhd/status/1575345301010976771
page_id= doc_2886 – page_title: https://www.twitter.com/u_me_reality/status/1575373775356547074
page_id= doc_2492 – page_title: https://www.twitter.com/GabeZZOZZ/status/1575468755760685057

Figure 7

page_id= doc_3996 – page_title: https://www.twitter.com/IrishMirror/status/1575154617620504576

page_id= doc_156 – page_title: https://www.twitter.com/rukigafm/status/1575901742398861312

page_id= doc_143 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575904025756696582

page_id= doc_1588 – page_title: https://www.twitter.com/AfricaTembelea/status/1575672516395163648

page_id= doc_1464 – page_title: https://www.twitter.com/reconnxx/status/1575706093849694208

page_id= doc_2938 – page_title: https://www.twitter.com/JenJJams/status/1575363112437313536

page_id= doc_1197 – page_title: https://www.twitter.com/nehakhanna_07/status/1575782059066691585

page_id= doc_470 – page_title: https://www.twitter.com/MrFukkew/status/1575852450577551361

page_id= doc_632 – page_title: https://www.twitter.com/ndtv/status/1575829177973907456

page_id= doc_418 – page_title: https://www.twitter.com/ndtv/status/1575860393284509696

page_id= doc_2637 – page_title: https://www.twitter.com/SGNewsAlerts/status/1575450879343448066

page_id= doc_2393 – page_title: https://www.twitter.com/NEWS_ALL_TIME/status/1575487364168028162

page_id= doc_3692 – page_title: https://www.twitter.com/AtharInanloo/status/1575189460786073612

page_id= doc_3078 – page_title: https://www.twitter.com/ghethwa/status/1575320030736261121

page_id= doc_1967 – page_title: https://www.twitter.com/MoRaY1959/status/1575601559756382208

page_id= doc_403 – page_title: https://www.twitter.com/UATV_en/status/1575863041547288579

page_id= doc_177 – page_title: https://www.twitter.com/MenorRondon/status/1575899796350476288

page_id= doc_179 – page_title: https://www.twitter.com/ttpenews/status/1575899616012214272

page_id= doc_247 – page_title: https://www.twitter.com/4CARDS_Ent/status/1575890966489071619

page_id= doc_585 – page_title: https://www.twitter.com/latestly/status/1575835227561553920

Figure 8

page_id= doc_3094 – page_title: https://www.twitter.com/normansolomon/status/1575314387388317697

page_id= doc_6 – page_title: https://www.twitter.com/NEWS_ALL_TIME/status/1575917759707299841

page_id= doc_2476 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575471945411432452

page_id= doc_2674 – page_title: https://www.twitter.com/knittingknots/status/1575441987028787206

page_id= doc_376 – page_title: https://www.twitter.com/knittingknots/status/1575867325923741696

page_id= doc_696 – page_title: https://www.twitter.com/anthony51483709/status/1575825041731641346

page_id= doc_3740 – page_title: https://www.twitter.com/DonalSmithCllr/status/1575184469447036928

page_id= doc_844 – page_title: https://www.twitter.com/everywhere_war/status/1575816493815980032

page_id= doc_627 – page_title: https://www.twitter.com/Daddyspeakez/status/1575829557248430081

page_id= doc_895 – page_title: https://www.twitter.com/history22nd/status/1575812892758114306

page_id= doc_939 – page_title: https://www.twitter.com/Ayaneth_/status/1575809447569788928

page_id= doc_124 – page_title: https://www.twitter.com/BlogUkraine/status/1575905959280402432

page_id= doc_122 – page_title: https://www.twitter.com/Pr0fM0riarty/status/1575906052192673793

page_id= doc_3110 – page_title: https://www.twitter.com/jul4enek/status/1575308105394667521

page_id= doc_3532 – page_title: https://www.twitter.com/WarriorsWhisper/status/1575206589287452672

page_id= doc_3654 – page_title: https://www.twitter.com/FCU_Ukraine/status/1575193215392657409

page_id= doc_1442 – page_title: https://www.twitter.com/FreeCiviliansUA/status/1575714852299243520

page_id= doc_465 – page_title: https://www.twitter.com/Magnifyingglas_/status/1575853606330585090

page_id= doc_3228 – page_title: https://www.twitter.com/2536luis/status/1575266793764986880

page_id= doc_3054 – page_title: https://www.twitter.com/thomaszickell/status/1575329510807027712

Figure 9

```
page_id= doc_1741 - page_title: https://www.twitter.com/EUFreeCitizen/status/1575641892888649728
page_id= doc_1117 - page_title: https://www.twitter.com/Drobdcr/status/1575791991522140162
page_id= doc_2611 - page_title: https://www.twitter.com/slavamakarov/status/1575454937705840641
page_id= doc_633 - page_title: https://www.twitter.com/nat_telepneva/status/1575829074429157376
page_id= doc_2316 - page_title: https://www.twitter.com/GranataLLC/status/1575510230460497920
page_id= doc_2554 - page_title: https://www.twitter.com/_Nex3_/status/1575460847584960513
page_id= doc_1260 - page_title: https://www.twitter.com/swapnilzs/status/1575766081092083712
page_id= doc_2478 - page_title: https://www.twitter.com/Pr0fM0riarty/status/1575471379994009604
page_id= doc_3356 - page_title: https://www.twitter.com/Dreadedfull/status/1575242098197995520
page_id= doc_741 - page_title: https://www.twitter.com/SGNewsAlerts/status/1575822679139299328
page_id= doc_2705 - page_title: https://www.twitter.com/KacoBlokland/status/1575434206288683008
page_id= doc_3760 - page_title: https://www.twitter.com/financetwitting/status/1575182277130452992
page_id= doc_1251 - page_title: https://www.twitter.com/pka71/status/1575767954469392386
page_id= doc_1253 - page_title: https://www.twitter.com/AGCsegreteria/status/1575767861364236289
page_id= doc_1254 - page_title: https://www.twitter.com/AlbaneseAl/status/1575767802417451008
page_id= doc_2548 - page_title: https://www.twitter.com/_Nex3_/status/1575461490190172161
page_id= doc_1446 - page_title: https://www.twitter.com/Pr0fM0riarty/status/1575712265118040064
page_id= doc_457 - page_title: https://www.twitter.com/Suspended_Acct/status/1575854478544474112
page_id= doc_3660 - page_title: https://www.twitter.com/WWEverythingYT/status/1575192706971992065
page_id= doc_3090 - page_title: https://www.twitter.com/Mickey17176/status/1575314795573792768
```

Figure 10

Comparison with transformer-based embeddings

Transformer-based embeddings, such as those from BERT or RoBERTa, bring several enhancements and challenges to the information retrieval process compared to traditional embeddings like Word2Vec. This is the main reason why all state of the art models use transformer-based embedding techniques.

Some of the benefits of transformer-based architectures are the following

1. **Context:**
   These models provide contextual embeddings, capturing the meaning of a word based on its surrounding context. This is particularly beneficial for short texts like tweets, where context is crucial for understanding the nuances.

2. **Rich Representation**:
   These models capture complex relationships and semantics within a sentence. They can represent the nuances of language more effectively than Word2Vec, which provides a fixed embedding for each word.

3. **Handling Out-of-vocabulary Words**:

These models can handle out-of-vocabulary words better than traditional embeddings. They have a subword tokenization scheme, enabling them to represent rare or unseen words by breaking them into subword units.

4. **Fine-tuning Possibility:**
   Fine-tuning the pre-trained models on domain-specific data can enhance their performance in a specific context. This flexibility is advantageous when dealing with tweets that often have unique characteristics and informal language. We could fine-tune an already existing embedding system over all the tweets of our dataset in order to improve the performance of the embedding.

All the aforementioned performance benefits come with a significant tradeoff in terms of computational expense.

1. **Computational Overhead:**
   These models are computationally more expensive than Word2Vec, both in terms of training and inference. Deploying transformer models for real-time information retrieval might be challenging in resource-constrained environments. So, it would take longer to set up the search engine since all the tweets should be embedded previously to gain efficiency. Then, only the query should be embedded at search time.

2. **Model Size:**
   These models are larger and more complex than Word2Vec, making them more difficult to be deployed to a production environment and to maintain. Furthermore they require much more computational resources so a higher cost in time to maintain them.

3. **Fine-tuning Requirement:**
   Despite the fact that a vanilla embedding model (not fine-tuned) could even work better than Word2Vec, a fine-tune over existing tweets in our case would be very beneficial as explained above. This process is not easy at all since it requires to process all the data, to modify some of the complex layers of the model to be re-trained.

In conclusion, while transformer-based embeddings offer superior contextual understanding and representation of nuances in short texts like tweets, their use comes with computational challenges and the need for careful consideration of the trade-offs. The decision to use BERT or RoBERTa over traditional embeddings should be based on the specific requirements of the information retrieval task, available resources, and the importance of contextual understanding in capturing the semantics of short texts.