A PROTOTYPE DEEP LEARNING PARAPHRASE IDENTIFICATION SERVICE FOR DISCOVERING INFORMATION CASCADES IN SOCIAL NETWORKS

Panagiotis Kasnesis[*], *Ryan Heartfield*[†]*, Lazaros Toumanidis*[*]*, Xing Liang*[†]*, George Loukas*[†]*,*
and Charalampos Patrikakis[*]

[*]{laztoum,pkasnesis,bpatr}@uniwa.gr; [†]{r.heartfield,x.liang,g.loukas}@gre.ac.uk

ABSTRACT

Identifying the provenance of information posted on social media and how this information may have changed over time can be very helpful in assessing its trustworthiness. Here, we introduce a novel mechanism for discovering "post-based" information cascades, including the earliest relevant post and how its information has evolved over subsequent posts. Our prototype leverages multiple innovations in the combination of dynamic data sub-sampling and multiple natural language processing and analysis techniques, benefiting from deep learning architectures. We evaluate its performance on EMTD, a dataset that we have generated from our private experimental instance of the decentralised social network Mastodon, as well as the benchmark Microsoft Research Paraphrase Corpus, reporting no errors in sub-sampling based on clustering, and an average accuracy of 92% and F1 score of 93% for paraphrase identification.

Index Terms— Information Cascade, Clustering, Deep Learning, Paraphrase Identification

1. INTRODUCTION

To assess the veracity of social network post information, it is useful to identify sources of the information, when it was first published and in which post, how similar the information is to other published sources, as well as how all related sources of information may have changed over time. To model this systematic flow of related information, we define this identification process as an information cascade, whereby the information cascade can be described as a two-step process. Firstly, a user encounters a piece of information and acts upon it by sharing it. The act of information sharing as a result of coming across a particular post creates a new cascade or further expands an existing one. Secondly, through the process of sharing, the user may adopt or support the position of the information, amplifying it; or rejects the information via direct intervention by correcting or commenting on it and in this way modifying it. In this paper, we introduce a mechanism for identifying information cascades which employs Paraphrase Identification (PI) based on deep learning as its core technological component. The mechanism facilitates the discovery of the source of information within the confines of a decentralised social network, which in our case is a private instance of Mastodon, the most popular decentralised social media platform. We refer to this as the EUNOMIA [1] Mastodon instance, named after the research project in which it has been set up. The information cascade mechanism also establishes cascade connectivity for related information. The aim is to help users assess themselves the trustworthiness of information they see on social media more efficiently and based on more information than is normally available. Note that in line with the scope of project EUNOMIA's approach, the information cascade mechanism works on social media posts where the users have provided their consent to be indexed by it. Contrary to the usual focus in information cascade research, which is on early prediction of whether a particular post will "go viral" [2, 3], our focus is on recognising cascades that have already occurred within the context of helping assess information trustworthiness, especially where the information has been modified across different posts and are not identical reshares.
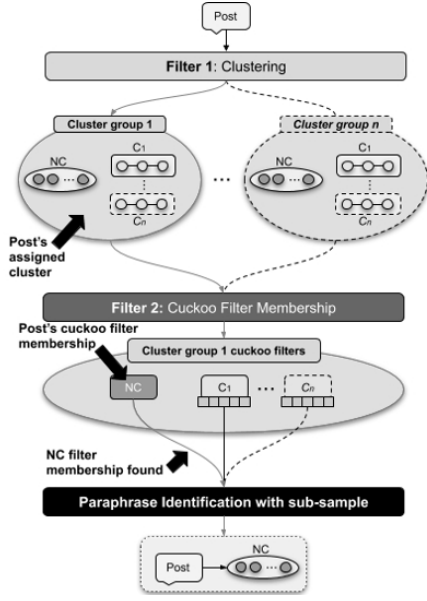
2. INFORMATION CASCADE MECHANISM

The *information cascade mechanism* functions as a series of linear-interacting system processes. There are four main stages to the creation or addition to an information cascade (Figure 1): 1) the extraction of post data and meta-data, 2) the sub-sampling of existing EUNOMIA-indexed posts for comparison analysis, 3) the comparison of semantic similarity between sub-sampled posts and the uploaded posts, and 4) The creation of a new information cascade, addition of a post to an existing information cascade, or no further action. In the next sections we will describe the functional behaviour of the *information cascade mechanism* and the processing steps for post ingesting, sub-sampling, PI, and cascade membership.

2.1. Cascade Post Subsampling Process

2.1.1. Feature Extraction

Within the information cascade mechanism, the text feature extraction service is responsible for converting post input

Fig. 1. Information Cascade mechanism overview (NC - None cascade post, C*n* - Cascade post)

(which consists of tokenised posts) into an array of float numbers that contain the post text's semantics. In particular, EUNOMIA exploits Transformer-based deep learning models [4] that have been pre-trained in an unsupervised manner by using a "masked language model" objective; it randomly masks some of the input tokens and aims to predict the token id of the masked word utilising only its context. E.g, in the sentence: *EUNOMIA is a European [Mask] action project*, the model must predict the id of the "innovation" token.

2.1.2. Post Clustering and Cuckoo Filtering

For post subsampling we leverage cluster group assignment alongside Cuckoo filters [5] as a highly efficient means of subsampling data probabilistically, because from a post content perspective it supports the grouping of semantically similar posts based on their content. Therefore, by using clustering techniques we can establish a confidence metric that grouped posts contain similar words or derived features (i.e., sentence embeddings) that make them suitable for filtering. In the first prototype we have explored multiple clustering techniques and feature vector extraction technologies which represent the state-of-the-art in open source literature, namely, hierarchical and k-means clustering and RoBERTa [6], BERT [7] and Universal Sentence Encoder (USE) [8] for feature extraction; all of which utilise unsupervised learning models with pretrained Transformer encoders. Experimental testing on EMTD is discussed in section 3.1.

2.2. Paraphrase Identification Classification Service

The PI classification service defines the similarity score between two posts and, as a consequence, if they are linked with each other in the information cascade. In the EUNOMIA cascade creation flow it comes after the filtering services (e.g., post subsampling) since it is computationally expensive service, particularly as the number of posts in a EUNOMIA instance grows. Thus, after subsampling the set of the possible similar posts (from the cluster cuckoo filter output), the input post (anchor post) is compared with each post in this subsample, producing a probabilistic output (range 0-1) for each pair of posts. For the first prototype, we have selected the BERT model [7], finetuned (using transfer learning) on the MRPC dataset (see section 3.2), utilising the saved model to perform a sentence pair classification task within a first-in first-out data pipeline for uploaded user posts.

After PI is executed, we select the post which has the highest similarity score with the anchor post. Where the similarity exceeds a predefined threshold (e.g., 0.8), it is considered to be the parent node and the anchor post its child (leaf) node. In this way, a cascade tree is constructed by adding leaf nodes to display how the information is linked together in the cascade, with time delta (i.e., post creation time) between linked posts determining the directionality of the link. In case the aforementioned criteria are not satisfied, the anchor post is not linked to any of the existing cascades but is considered to be a contestant root node. The information cascade mechanism utilises JSON as a portable data structure for transporting data and it is used to represent post and information cascade objects. Figure 2 provides an example of the post meta-data information containing the cascade links (represented in light grey).

```
{ "id": "string",
  "source_id": "string",
  "source_url": "string",
  "account_id": "string",
  "created_at": "2020-03-26T21:23:55.379Z",
  "post_text": "string",
  "mentions": ["string"],
  "votes_trust": 0,
  "votes_notrust": 0,
  "votes_neutral": 0,
  "your_trust": {
    "trust": true,
    "comment": "string"},
  "cascade_id": "string",
  "sentiment_class": 0,
  "sentiment_confidence": 0,
  "subjectivity_score": 0,
  "cascade_links": {
    "parent": {
      "post_id": "string",
      "time_delta": 0,
      "paraphrase_similarity": 0,
      "link_similarity": 0,
      "img_similarity": 0,
      "video_similarity": 0,
      "hashtag_similarity": 0 },
    "children": [
      { "post_id": "string",
        "time_delta": 0,
        "paraphrase_similarity": 0,
        "link_similarity": 0,
        "img_similarity": 0,
        "video_similarity": 0,
        "hashtag_similarity": 0 }]}
}
```

Fig. 2. EUNOMIA-enriched Post JSON Object

3. EXPERIMENTAL PROTOTYPE ANALYSIS

3.1. Clustering - Sentence Embedding results

We evaluated the performance of the clustering methods with respect to that of the sentence embeddings using the posts contained in the EMTD. In particular, we tested hierarchical and kMeans clustering algorithms combined with three state of the art sentence embedding extractors (RoBERTa, BERT [7] and USE [8]. The evaluation metric was recall, i.e. if the same paraphrased posts belong to the same cluster. RoBERTa combined with hierarchical clustering (Figure 3) obtained the best results (100% accuracy with EMTD).
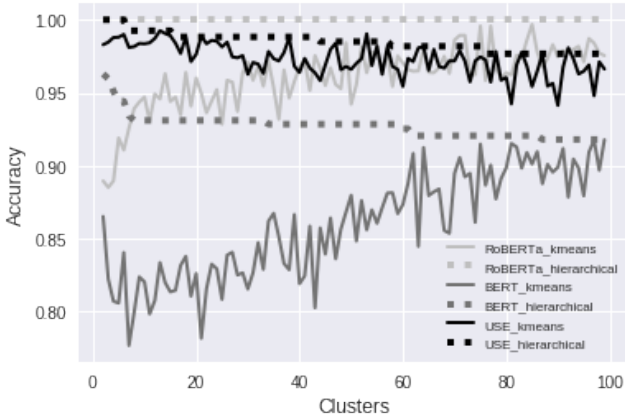

Fig. 3. Clustering accuracy.

3.2. Paraphrase Identification results

The experiments were executed on a computer workstation equipped with a NVIDIA GTX 1080 Ti GPU featuring 11 gigabytes RAM, 3584 CUDA cores and a bandwidth of 484 GB/s. We used Python as programming language, and specifically the Numpy library for matrix multiplication, Re library for text preprocessing (i.e., regular expression operations) and Pytorch library for retraining and evaluating the BERT model. In order to accelerate the tensor multiplications, we used the CUDA Toolkit in support with the cuDNN, which is the NVIDIA GPU-accelerated library for deep neural networks.

To evaluate the performance of the prototype PI service, we trained the model using MRPC (Microsoft Research Paraphrase Corpus) [9], included in the GLUE (General Language Understanding Evaluation) benchmark [10], and the EMTD (EUNOMIA Mastodon Testbed Dataset) to test its performance. MRPC is a corpus of sentence pairs included in online news sources, annotated by humans to define whether the sentences in the pair are semantically equivalent; it is imbalanced (68% positive, 32% negative). EMTD[1] is a specifi-

---
[1]https://github.com/consert/emtd

Table 1. PI model performance

Dataset	Accuracy (%)	F1 score (%)
MRPC	85.35	89.81
EMTD	97.78	96.72

cally crafted dataset of posts on an isolated Mastodon testbed, using dummy Mastodon user accounts for emulating user posts and information cascades. Compared to the MRPC dataset size of 5801 sentence pairs, EMTD consists of only 616 posts where we have also generated paraphrase sets consisting of 3-4 posts to emulate information cascades (containing 78 in total); which the MRPC dataset does not contain. The EMTD dataset was used for model testing only.

The results in Table 1 illustrate the performance of the fine-tuned BERT model for the PI task, while Figure 4 displays the confusion matrix for the EMTD. It should be noted that the model is not very sensitive, tending to predict paraphrased samples as irrelevant (i.e. it has 20 false negatives) especially in cases where post contains uncommon words (e.g., disbursal instead of spending) or numbers instead of words (e.g., 10 instead of ten).
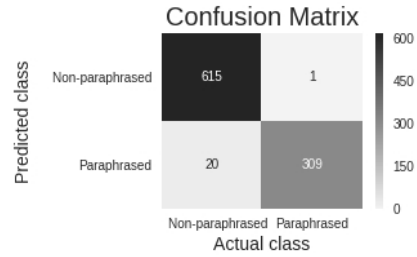

Fig. 4. Confusion Matrix displaying the results of the PI on the EMTD dataset.

In Figure 5, the results illustrate the computation time for the paraphrase identification process, using posts from the EUNOMIA Mastodon instance. The results clearly highlight that by employing a cluster-based subsampling methodology is significantly more efficient than a greedy method of comparing a post against the global set of posts.

3.3. Decentralised Network Cascade Detection Prototype

In terms of integration, the information cascade discovery has been deployed to the Digital Companion application [11], which is the EUNOMIA client for accessing content published in decentralised social networks. It allows user-friendly interaction and visualisation of the cascades, as well as enabling the users to have access to other type of information, called EUNOMIA indicators (e.g., the sentiment of post) or even vote whether a post is trustworthy or not. Figure 6
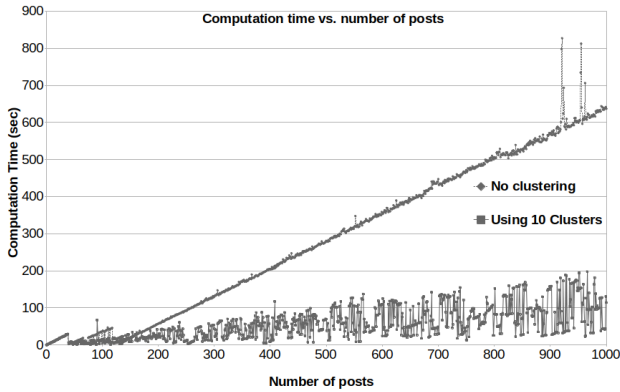
Fig. 5. Comparison of computation time with and without cluster-subsampling in PI
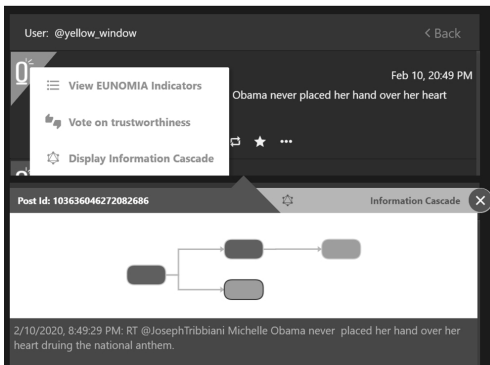
Fig. 6. Screenshot displaying an information cascade included in the EUNOMIA Mastodon instance.

presents part of the Digital Companion's user interface, displaying the tree-based representation of a cascade.

4. CONCLUSIONS

In this paper we have presented a prototype information cascade mechanism utilised to discover information cascades in decentralised social networks, such as Mastodon. The information cascade mechanism can play a key role in helping a user assess the trustworthiness of information posted on social media, by identifying the earliest appearance of this information and how it may have evolved over time. The natural language processing and analysis mechanisms we have developed and integrated into the prototype information cascade mechanism have reported excellent experimental performance; tested using on posts recorded in two datasets. Future development will expand the mechanism by taking into account multimedia similarity in order to provide a more diverse information cascade creation service. Moreover, the training and deployment of an explainable deep learning algorithm for textual entailment annotations amongst the posts included in a cascade will also be considered.

6. REFERENCES

[1] "Eunomia eu h2020 project," https://eunomia.social/, (accessed: 04.20.2020).

[2] Karthik Subbian, B Aditya Prakash, and Lada Adamic, "Detecting large reshare cascades in social networks," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 597–605.

[3] Wei Xie, Feida Zhu, Jing Xiao, and Jianzong Wang, "Social network monitoring for bursty cascade detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 4, pp. 1–24, 2018.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[5] Bin Fan, David G. Andersen, Michael Kaminsky, and Michael Mitzenmacher, "Cuckoo filter: Practically better than bloom," in *CoNEXT '14*, 2014.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[8] Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Universal sentence encoder," *ArXiv*, vol. abs/1803.11175, 2018.

[9] William B. Dolan and Chris Brockett, "Automatically constructing a corpus of sentential paraphrases," in *IWP@IJCNLP*, 2005.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *BlackboxNLP@EMNLP*, 2018.

[11] Lazaros Toumanidis, Ryan Heartfield, Panagiotis Kasnesis, George Loukas, and Charalampos Z. Patrikakis, "A prototype framework for assessing information provenance in decentralised social media: The eunomia concept," in *e-Democracy*, 2019.