October 2020

Protecting Authentic Entities On Social Media From Impostors

Anonymous

Protecting Authentic Entities On Social Media From Impostors

ABSTRACT

Social media and other online platforms are frequently confronted with the problem of impostor accounts, e.g., individuals posing to be someone else on the platform. A conventional approach to identifying impersonators is to use a seed-set of verified and protected entities to perform a search in the social-media graph to discover other entities that are similar enough to a verified entity to potentially be impersonators of the verified entity. However, many platforms often do not have a list of verified entities available a priori. This disclosure describes the use of machine-learning techniques to detect impostors on social media and other online platforms in scenarios where a verified seed-list of authentic entities is unavailable. Additionally, the techniques discover new verified seeds, e.g., small-scale businesses, local organizations, etc. that can then be placed on the protection radar.

KEYWORDS

- Social media
- Authentic entity
- Impersonator account
- Fake account
- Impostor
- Machine learning
- Verified entity
- Seedless detection

BACKGROUND

Social media and other online platforms are frequently confronted with the problem of impostor accounts, e.g., individuals posing to be someone else on the platform. The integrity of such platforms depends on their ability to differentiate authentic entities from impostors and to protect authentic entities from impostors. A conventional approach to identifying impersonators and fake platform-presence is to use a set of verified and protected entities, known as seeds, and to perform a search in the social media graph to discover other entities that are similar enough to a verified entity to potentially be impersonators to the verified entity. However, many platforms often do not have a list of verified entities available a priori.

A related problem is that some entities on social-media platforms may not claim to represent a certain identity, but are nevertheless closely related to that identity. For example, a movie star may have an authentic but unverified page, while also having several independent fan pages that are run by other people. Such fan pages are not strictly impostors. Yet, given the collection of fan pages, it is difficult to identify the movie star's true page.

In an extreme case, all accounts that pertain to an entity on a platform may be impostors, e.g., the authentic entity may not be present on the platform. In an opposite extreme case, there can be *multiple* authentic entities with the same name but different ownership, e.g., news channels with the same or nearly the same name (and logo) operating in different countries.

DESCRIPTION

This disclosure describes techniques to detect impostors and fake presence on social media and other online platforms. The techniques discover new verified seeds, e.g., small-scale businesses, local organizations, etc., that can then be placed on the protection radar. The techniques enable seedless detection of authentic entities, e.g., work in use-cases where a

verified seed-list of authentic entities is unavailable. Machine learning models are utilized to identify authentic entities and distinguish them from impersonating and/or fake entities.

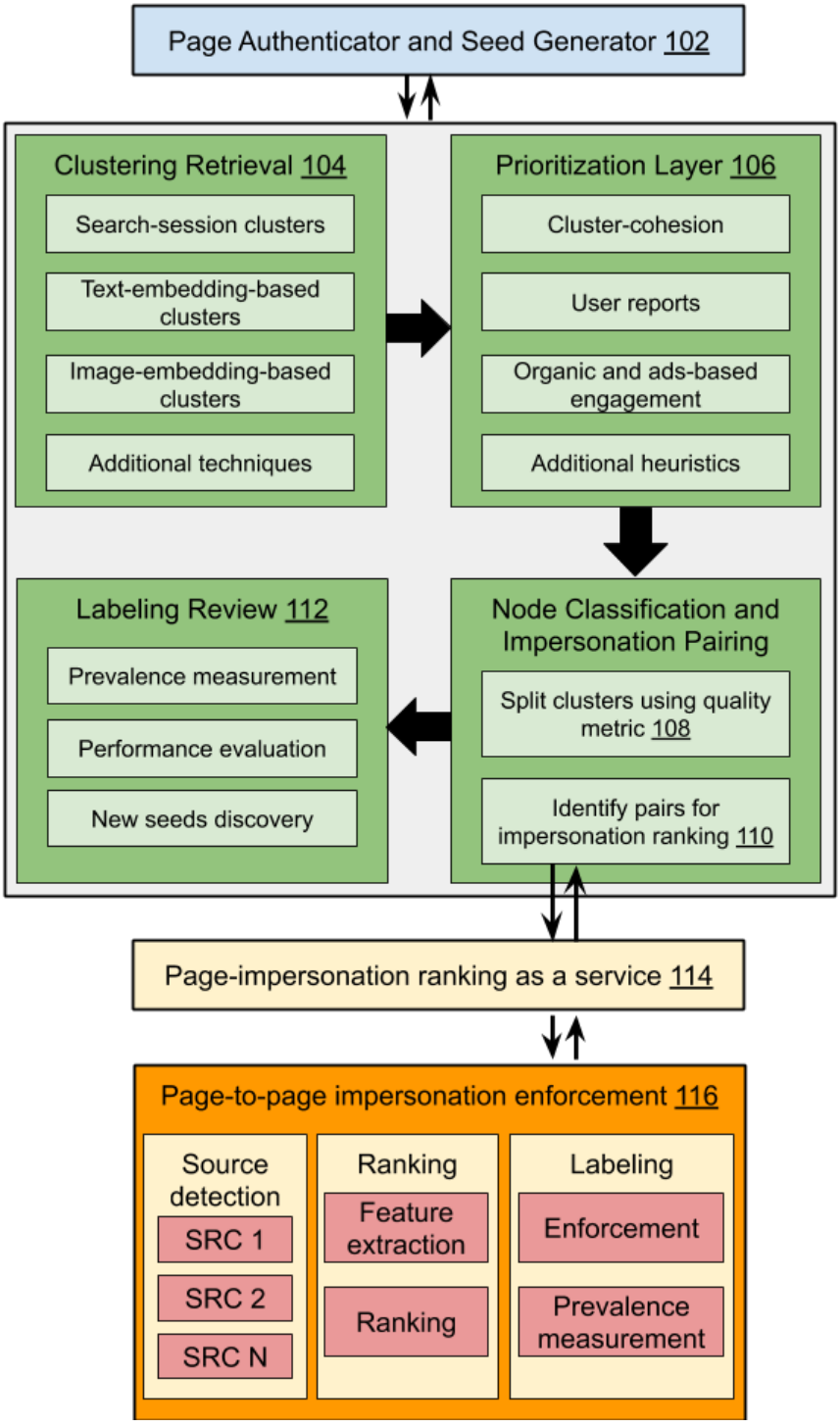

Fig. 1: Protecting authentic entities on social media from impostors

Fig. 1 illustrates an example detection framework for protecting authentic entities on social media (or other platforms) from impostors. The framework of Fig. 1 analyzes pages on the social media platform; other types of account-related content can be utilized in a similar manner. The framework includes the following components (layers):

Page authenticator and seed generator (102)

The page authenticator analyzes signals from pages of accounts on the platform to develop an initial estimate of the authenticity of the page. Example signals can include quality of page content, cohesiveness of page content, e.g., topic-specificity, deviation from the main topic, semantic alignment between text and images on the page, historical policy violations, user feedback and sentiment signals, etc. A page with a strong authenticity signal can potentially serve as a seed.

Source-clustering and aggregation layer (104)

Based on various attributes of entities, e.g., text or image-based n-grams found on the pages or accounts of the entities, content and semantic information of the pages, interactions of the accounts with their followers, hashtags on the page, search sessions on the platform, connectivity of the entity to other entities, embedding representations of each of the attributes, etc., the universe of entities is (even when no seed entities are known) clustered using machine-learning models into relatively large clusters. For example, entities that represent the same brand, business, or organization are clustered together. Different clustering methods are used to maintain high recall, ensuring that all true-positive detections are aggregated together, even at the cost of potentially including some false positives.

Prioritization layer (106)

The prioritization layer improves computational efficiency by enabling re-ranking after the source-clustering and aggregation layer. This is especially important in applications where the volume of the entities is very high, e.g., platforms that have a very large number of users, business, or page accounts. Pairwise comparison of a large number of entities, e.g., numbering in the hundreds of millions, to test the authenticity of one member of the pair against another is computationally infeasible. By discarding less-cohesive clusters, the prioritization layer enables reduction of the number of clusters that are processed further to detect impersonating entities. Prioritization can be based on cluster-cohesiveness, user reports (e.g., of policy violations by certain accounts), organic and/or advertising-based user engagement, and other heuristics.

Cluster-splitting layer (108)

A quality metric is defined that scores the authenticity of entities in the social-media graph. Effectively, the quality metric serves as a proxy for authenticity. Each prioritized cluster is split by quality score into potentially authentic and inauthentic sets of nodes. A cross-join of all possible pairs from one set to the other is performed to evaluate the possible "impersonation" edges within this bipartite graph for each cluster. Although the quality metric in standalone form does not by itself enable the identification of notable entities on the platform, it does enable the identification of high-quality, e.g., potentially authentic, nodes in a cluster.

Layer for pairwise evaluation of impersonation-similarity (110)

This layer evaluates a given pair of nodes for impersonation through the similarity of their content. The layer provides at its output ranked pairs for possible impersonating cases. The output from the previous layer, e.g., node pairs from cluster splits, are ingested by this layer for

the purposes of impersonation evaluation. This layer effectively filters out false positives from clusters. The rankings generated by this layer can be accessed as a service (114).

The output of the pairwise-evaluation layer can be used to identify clusters that contain a high number of impostors; these are accounts (e.g., of brands, users,, or organizations) of substantial value, worthy of attracting impostors. The techniques thus enable the discovery of potentially new authentic seeds. The output of the pairwise-evaluation layer can also be used to identify multiple authentic entities, such as brands with the same name but different ownerships operating in different countries.

Labeling review layer (112)

Once the authentic seeds are identified using machine learning models, their impostors, and their non-impostors, entity samples from different score-bucket regions are subjected to human review. The purpose of such human labeling review is to evaluate the performance of the machine learning models, e.g., to map the confidence score generated by the model to a true-positive or false-positive rate; to further train the machine learning model by providing feedback on performance; to estimate the prevalence of authentic versus impostor entities; to assist in the discovery of new seeds; etc.

Layer for impersonation enforcement (116)

Enforcement is a procedure by which identified impersonators for protected seeds are notified and their accounts or pages taken down from the platform. Seed-based enforcement includes components such as source detection, where high-recall impostors are formed around targeted (known) seeds; ranking, where potential impostors to a seed are ranked by impersonation probability in a pairwise fashion; labeling, where the most probable impostors are sent for human review, take-down, and prevalence measurement. The enforcement procedure

results in the generation of training data for machine-learning based impostor clustering and seed generation. For example, a verified seed, e.g., a well-known brand, and the collection of its impostors as identified by the enforcement procedure, can serve as training data.

Enforcement can take a variety of forms. For example, an outright case of impersonation is treated by taking down the impersonating account. A benign case of impersonation, e.g., a fan page of a movie star, or a help/support group for a brand or product, can be treated by instructing the benign impersonator to post a notice clarifying their relationship to the movie star or brand.

In this manner, the techniques described herein enable the detection of impostors on a social media or other online platform in scenarios where a verified seed-set of entities is unavailable. The techniques also enable the discovery of potentially new authentic entities on the platform, where simple quality-based scoring prediction may be insufficient.

CONCLUSION

This disclosure describes the use of machine-learning techniques to detect impostors on social media and other online platforms in scenarios where a verified seed-list of authentic entities is unavailable. Additionally, the techniques discover new verified seeds, e.g., small-scale businesses, local organizations, etc. that can then be placed on the protection radar.