Existing methods research

Phan Lê Việt Hùng



# Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model for Instagram

Cyberbullying is a serious and growing problem on social networks. Traditional supervised learning methods for detecting cyberbullying require large amounts of labeled data. Collecting and labeling this data is time-consuming, expensive, and may not keep up with changing forms of cyberbullying.

However, unsupervised learning methods also have some challenges:

* **Challenges in time modeling:** Because cyberbullying is often a repetitive behavior, temporal analysis provides more detailed information than text-only methods, which treat each comment as a separate event. Therefore, we need to simultaneously model both the time factor and online bullying detection, so that these two tasks support each other.
* **Challenges in modeling multi-layer structures:** Interactive sessions on social networks inherently have a multi-layered structure, where words make up comments and comments make up a session. Previous studies show that modeling this multilayer structure is useful for learning high-quality representations. Additionally, because the meaning of words and comments depends heavily on context, it is necessary to accurately model their sequential structure to identify relevant factors.
* **Limitations of cluster methods:** Available cluster algorithms (e.g. k-means) depend largely on the quality of the input data. Social network data is famous for being noisy, sparse, and multi-dimensional. Performing dimensionality reduction on input data still suffers from the disadvantage of decoupled training, meaning that representation and cluster learning are performed separately.

The author proposed Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model (UCD) method based on the time-based Gaussian mixture model.

The time-varying Gaussian mixture model is a statistical model that can learn probability distributions from data over time. In this model, each probability distribution represents a different type of behavior. The model can learn to distinguish patterns of cyberbullying behavior from normal patterns of behavior, as well as detect new instances of cyberbullying by identifying data that does not fit any of the probability distributions.

UCD consists of two main components: a representation learning network, which learns the compact multi-modal representations of a session; and a multi-task learning network, which predicts whether or not a session contains bullying behaviors while modeling the temporal dynamics of all comments. Specifically, the representation learning network uses a Hierarchical Attention Network (HAN) [47] for textual features and a Graph Auto-Encoder (GAE) [23] for user and network features. The multitask learning network then takes the multi-modal representations (e.g., text, user, and social network) as input to estimate the bullying likelihood using a time-informed Gaussian Mixture Model (GMM). The two UCD components are jointly optimized to boost their learning effectiveness.

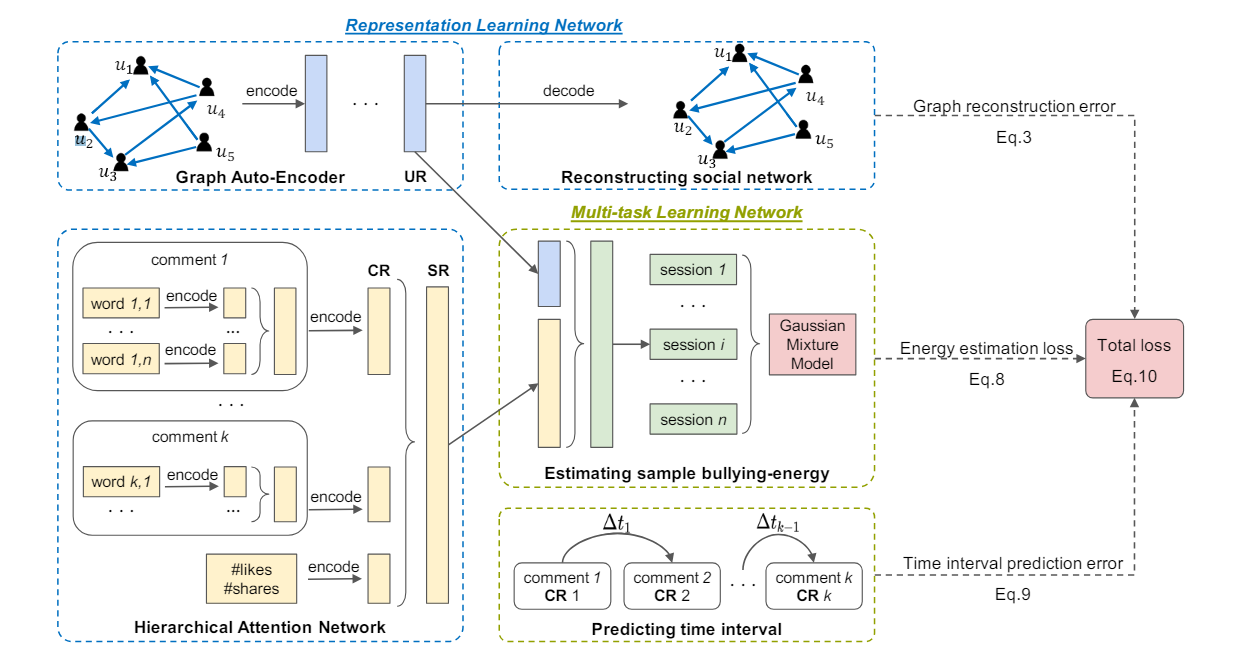
Advantage:

* No labeled data is required.
* Be able to adapt to changing types of cyberbullying over time.

Disadvantages:

* Normal behavioral patterns can be confused with cyberbullying.
* Affected by data quality.

## Representation Learning Network



**Graph Auto-Encoder (GAE):** Encodes user characteristics into User Representation (UR) into a lower dimensional representation, but still retain important information about that user.

**Hierarchical Attention Network (HAN):** Encode comments into Comment Representation (CR) and combine with react encodings into Social Representations. Additionally, it models the two main levels of social media sessions (sequences of words and comments) and at each level, the model captures the long-term dependencies and integrates mechanisms to differentiate the importance of specific words and comments based on their context.

## Multi-task Learning Network

**Sample bullying energy estimation task**

Use Gaussian Mixture Model (GMM) to estimate the density estimator. GMM is a multi-modal distribution model that is combined from many unimodal Gaussian distributions.

Suppose the number of mixture components in GMM is K and the implicit representation of an interactive session on a social network is s. First, the multi-task learning network predicts which mixture component s belongs to. Then, based on the above prediction results, the parameters of GMM are estimated so that the bullying level of session s (energy estimation) can be calculated. Specifically, s is fed into a multi-layer network (MLN) parameterized by θ\_m. The output of this MLN is a probability vector.

**The time interval between comments prediction task**

Information about the time interval between comments can be useful for distinguishing between instances of bullying and normal interactions. For example, in cases of bullying, users may comment continuously at a high frequency, while normal interactions may have longer intervals between comments.

# Detection Of Types Cyber-bullying Using Fuzzy C-means Clustering And XGBoost Ensemble Algorithm

In this study, a neural network model (XGB\_CTD) that will predict which type of bullying the users may be exposed to, through a dataset gained by the cyberbullying scale applied to young internet users is formulated. Extreme Gradient Boosting (XGboost) algorithm, one of the ensemble learning methods is used in this method. While this model contains 13 input parameters taken from the scale, there exists one output parameter classified as one of the 9 outputs. The reliability of the data set obtained through the survey is confirmed by statistical methods. The data set has been fragmented with Fuzzy C-Means (FCM) which is one of the fuzzy clustering algorithms. Hyper-parameters for the maximum efficiency of the model training have been defined as model, learning, and boosting methods. Independent variables in the data set have been scaled through standard normalization. As a result, the model has yielded a % 91,75 accuracy rate in the prediction of the classification as 9 different cyberbullying types. The same data set has been trained by different machine learning algorithms. It is seen that the proposed model has reached the highest accuracy when compared to the conventional machine learning algorithms.

## Ensemble learning with FCM and XGBoost

## 

* **Fuzzy C-Means (FCM)** giúp chia tập dữ liệu thành các cụm mờ, trong đó một điểm dữ liệu có thể thuộc về nhiều cụm với các mức độ khác nhau. Điều này giúp xử lý các trường hợp dữ liệu không rõ ràng hoặc có tính chất trung gian giữa các cụm khác nhau.
* **XGBoost** được áp dụng sau đó để tăng cường khả năng phân loại của các cụm đã được xác định bằng cách tối ưu hóa các cây quyết định và xử lý các mối quan hệ phi tuyến giữa các biến.

# REFERENCES

1. [Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model](https://avalanchesiqi.github.io/files/cikm2020cyberbullying.pdf)
2. [DETECTION OF TYPES CYBER-BULLYING USING FUZZY C-MEANS CLUSTERING AND XGBOOST ENSEMBLE ALGORITHM](https://crj.cit.edu.al/wp-content/uploads/2014/10/DETECTION-OF-TYPES-CYBER-BULLYING-USING-FUZZY-C-MEANS-CLUSTERING-AND-XGBOOST-ENSEMBLE-ALGORITHM.pdf)