Dear Sir or Madam,

Much appreciate for your invaluable suggestions. My answers are as follows,

Comment 1: The paper is fairly well written (though a spell-check session is necessary, due to the presence of some typos) and interesting.
Response: The updated manuscript and full version have been professionally edited and proofread by the language specialists from American Journal Experts. The released full version has not been submitted to anywhere else for publication. This full version can be accessed at https://github.com/mrspider520/gated_fusion_network/blob/master/paper/paper.pdf .

Comment 2: The experimental evaluation of the technique shows a slight improve over other state of the art methods (CNN reaches F1 .9 vs the paper's technique which is .92). The paper does not mention the configuration effort of the other techniques tested, therefore one could speculate that the comparison with other models is not necessarily fair.
Response: Due to the limited pages, we did not cover all of our comparison work. However, we would like to add more details in the revised version. In addition, we have released a full version of our work which has not been submitted to anywhere else for publication. This full version can be accessed at https://github.com/mrspider520/gated_fusion_network/blob/master/paper/paper.pdf. It provides extra experiments about model tuning and model comparison with different data sets in terms of scale and imbalance degree. The proposed framework shows the advantage to deal with larger and more imbalanced data with multimodal interactions. We also released the data and source code with this publication, in which the details about model training and tuning were presented.

Comment 3: Finally, I have concerns about the actual importance and relevance of the tackled problem. What is the actual impact of this problem? To what extent the gate/interaction mechanisms developed can be reused in other domains? Can't authors raise the problem to the identification of "personal" documents (including, for example, Facebook profiles) and then mention that one possible instance of the problem is the identification of faculty homepages?
Response: Recognizing faculty homepages is a good example of multimodal classification problem with interactive feature modes and imbalance data. Common strategies in previous studies have been either to concatenate features of various information sources into a compound vector or to feed the features separately into different classifiers, which are then assembled into a stronger classifier for the final decision. Both approaches inevitably ignore the interactions among different features. The layout feature set consists of three tags, that is, <*title*>, <*p*>, and <*footer*>. Each tag contains some textual information. The words embedded in the tag <*title*> are more important than those within the tag <*footer*>. Such interlinks are ignored, however, if the tag and text features are concatenated into a compound vector. In

addition, imbalanced data may cause deterioration of the classification performance, because the model pays less attention to the minority classes.

The proposed multimodal generative and fusion framework is able to be generalizable to many other multimodal learning problems with class-imbalanced data and interactive feature modes. A good example is the prediction of media-aware stock movements, in which the market information space consists of several interactive modes, including transaction data, news articles, and investors' mood [1]. In bear markets, most stocks have downward pressure, that is, most samples are negatives that lead to a serious class-imbalanced challenge. However, the effectiveness in related fields remains to be explored in the near future.

[1] Qing Li, Yuanzhu Chen, Li Ling Jiang, Ping Li, and Hsinchun Chen. A tensor-based information framework for predicting the stock market. ACM Transactions on Information Systems (TOIS), 34(2):11, 2016.