动机：

1）目的：使用基本且有缺陷的作者身份验证器（BAFF）来完成作者识别任务，并分析同一个模型在不同数据集上的性能偏差原因。

2）思路：用TFIDF加权3-gram模型加一些额外功能来扩展模型，分析造成结果偏差的原因并提供解决方法。

解决思路：

1）框架：{p, q, r}

构建一个由p，q和七个度量值构成的7维向量组成的三元字符向量。

2）模型细节：

七个度量值分别为：

1. 余弦相似度（TF加权）

2. 余弦相似度（TFIDF加权）

3. Kullback-Leibler散度（KLD）

4. 偏斜散度（skew-balanced KLD）

5. Jensen-Shannon散度

6. Hellinger距离

7. 平均对数句子长度差异（PAN参与者经常使用的特征）

七个度量值组成向量后需要对每个变量进行归一化处理，使每个值都处于[0,1]集合中。

偏差分析：

模型偏差：

B1：语料库相关特征。训练时选取与语料库相关的特征，会在测试时引起性能的严重下降。因此，应该避免与语料库相关的特征。

B2：特征缩放。对特征值进行归一化处理时，会使特征偏向于语料库的细节。

数据偏差：

B3：纯文本异质性。数据集文本中有许多部分不能表明作者风格，而是院子纯文本转换器或人工转录。另外有标点符号混用如ASCII何Unicode符号混用。

B4：人口同质性。将较小规模的数据扩充成较大数据时，部分文本会重复使用，导致作者的风格被低估或高度。

B5：意外文本重叠。案例之间的重叠（如重复的短语、重复的文本块）可能会使学习算法误入歧途。

评估偏差：

B6：测试合并。在测试时，作者身份验证器通常可以访问整个数据集；而在实际工作中，案件相互独立，验证器一次只处理一个案例，而不是相互参考。

实验：

1）数据集：PAN15测试数据集，PAN14小说数据集，PAN14论文数据集

2）baseline：PAN15和PAN14的参赛选手投稿的top5模型

3）创建数据集：the Webis Authorship Verification Corpus 2019

总结：在共享任务中，有时基本方法会胜过更复杂的机器学习方法，尤其是数据小的时候。进行作者识别时需要格外小心特征选择，应选择风格特征而不是主题特征。

其他：