

### 垃圾邮件一条龙: 从伪造到分类

网络信息内容安全作业二

刘迅

2024/06/19





### 目录

1 垃圾邮件内容分类实验

► 垃圾邮件内容分类实验 实验总览 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



### 目录

1 垃圾邮件内容分类实验

► 垃圾邮件内容分类实验 实验总览 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



### 实验目标

#### 1 垃圾邮件内容分类实验

- 本次实验要求完成 24000 封真实生产环境下的垃圾邮件日志的分类工作。
- 要求分为五类,但类别的定义未提供,需要自己设定。



图:数据集中的垃圾邮件示例



1. 基于 LLM Chat 的分类:



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。
  - 1.2 使用国内主流的大语言模型(Kimi, Qwen, ERNIE等)对邮件直接进行分类。



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。
  - 1.2 使用国内主流的大语言模型(Kimi, Qwen, ERNIE等)对邮件直接进行分类。
- 2. 基于 LM Embedding 的分类:



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。
  - 1.2 使用国内主流的大语言模型(Kimi, Qwen, ERNIE等)对邮件直接进行分类。
- 2. 基于 LM Embedding 的分类:
  - 2.1 使用预训练模型得到邮件内容的嵌入向量表示(Embedding)。



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。
  - 1.2 使用国内主流的大语言模型(Kimi, Qwen, ERNIE等)对邮件直接进行分类。
- 2. 基于 LM Embedding 的分类:
  - 2.1 使用预训练模型得到邮件内容的嵌入向量表示(Embedding)。
  - 2.2 使用 PCA,TSNE 方法对 Embedding 降维。



- 1. 基于 LLM Chat 的分类:
  - 1.1 根据提示工程学设计少样本系统指令。
  - 1.2 使用国内主流的大语言模型(Kimi, Qwen, ERNIE等)对邮件直接进行分类。
- 2. 基于 LM Embedding 的分类:
  - 2.1 使用预训练模型得到邮件内容的嵌入向量表示(Embedding)。
  - 2.2 使用 PCA,TSNE 方法对 Embedding 降维。
  - 2.3 观察是否能够聚类。



# 主要结果

#### 1 垃圾邮件内容分类实验

Kimi:推广營销	Qwen : 推广营销
Kimi:中文色情赌博	Qwen: 中文色情賭博
Kimi: 欺诈钓鱼	Qwen:欺诈钓鱼
Kimi:灰色产业	Qwen:灰色产业
Kimi:英文约会交友	Qwen: 英文约会交友

图: Kimi 与 Qwen 分类结果的桑基图交叉验证



#### 目录

1 垃圾邮件内容分类实验

▶ 垃圾邮件内容分类实验

头验忌觅

基于 LLM Chat 的垃圾邮件分类

基于 LM Embedding 的垃圾邮件分类 个例研究

▶ 发件人身份伪造实验

前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



1 垃圾邮件内容分类实验

#### 本次实验共使用三种国内主流的大语言模型:

- 1. Kimi (月之暗面) 版本为 moonshot-v1-8k。
- 2. Qwen(阿里)版本为 qwen-long。
- 3. ERNIE(百度)版本为 ernie\_speed。

#### 预设的五类垃圾邮件分别为:

1. 推广营销: 折扣优惠、活动邀请、培训班信息、红包奖励、学术会议推广等。



1 垃圾邮件内容分类实验

#### 本次实验共使用三种国内主流的大语言模型:

- 1. Kimi (月之暗面) 版本为 moonshot-v1-8k。
- 2. Qwen(阿里)版本为 qwen-long。
- 3. ERNIE(百度)版本为 ernie\_speed。

- 1. 推广营销: 折扣优惠、活动邀请、培训班信息、红包奖励、学术会议推广等。
- 2. 中文色情赌博:成人内容、不适当的语言或图像链接。



1 垃圾邮件内容分类实验

#### 本次实验共使用三种国内主流的大语言模型:

- 1. Kimi(月之暗面)版本为 moonshot-v1-8k。
- 2. Qwen(阿里)版本为 qwen-long。
- 3. ERNIE(百度)版本为 ernie\_speed。

- 1. 推广营销: 折扣优惠、活动邀请、培训班信息、红包奖励、学术会议推广等。
- 2. 中文色情赌博:成人内容、不适当的语言或图像链接。
- 3. 欺诈钓鱼: 试图通过伪装成可信来源诱骗收件人提供个人信息或进行恐吓。



1 垃圾邮件内容分类实验

#### 本次实验共使用三种国内主流的大语言模型:

- 1. Kimi(月之暗面)版本为 moonshot-v1-8k。
- 2. Qwen(阿里)版本为 qwen-long。
- 3. ERNIE(百度)版本为 ernie\_speed。

- 1. 推广营销: 折扣优惠、活动邀请、培训班信息、红包奖励、学术会议推广等。
- 2. 中文色情赌博:成人内容、不适当的语言或图像链接。
- 3. 欺诈钓鱼: 试图通过伪装成可信来源诱骗收件人提供个人信息或进行恐吓。
- 4. 灰色产业:包含提供发票、做账报销服务等内容。



1 垃圾邮件内容分类实验

#### 本次实验共使用三种国内主流的大语言模型:

- 1. Kimi(月之暗面)版本为 moonshot-v1-8k。
- 2. Qwen(阿里)版本为 qwen-long。
- 3. ERNIE(百度)版本为 ernie\_speed。

- 1. 推广营销: 折扣优惠、活动邀请、培训班信息、红包奖励、学术会议推广等。
- 2. 中文色情赌博:成人内容、不适当的语言或图像链接。
- 3. 欺诈钓鱼: 试图通过伪装成可信来源诱骗收件人提供个人信息或进行恐吓。
- 4. 灰色产业:包含提供发票、做账报销服务等内容。
- 5. 英文约会交友:包含英文的约会邀请、发照片、链接等。



### Few shot system messages

1 垃圾邮件内容分类实验

```
### 输入输出格式
基于以上系统提示、每次输入和输出的格式可以被设计如下
#### 输入1
感谢您一直以来对我们产品的支持,特此邀请您参加我们的VIP客户专享打折活动。点击此链接了解更多详情。
#### 输出1
#### 输入2
您的账户存在异常活动,请立即点击此链接验证您的身份,以确保您的账户安全。
#### 輸出2
散诈钓鱼
#### 输入3
第10周(2019年10月28日-2019年11月3日),人工智能学院开设科学前沿讲座安排如下、讲座名称讲座人讲座时间讲座地点以用户为中心的
#### 输出3
```

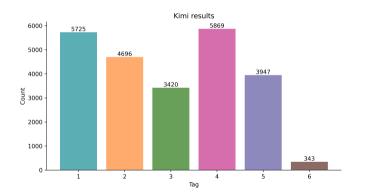
图: 通过提示工程学设计的少样本系统提示



# 分类结果: Kimi

1 垃圾邮件内容分类实验

Kimi 开发者: 月之暗面, 国内初创公司。



#### 标签含义

1:'推广营销'

2:'中文色情赌博'

3:' 欺诈钓鱼'

4:' 灰色产业'

5:'英文约会交友'

6:'高风险未判定'

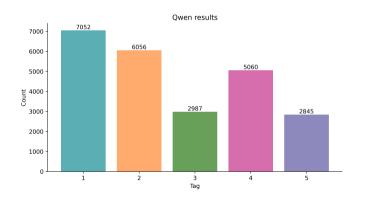
图: Kimi 分类结果



### 分类结果: Qwen

1 垃圾邮件内容分类实验

Qwen 开发者: 阿里巴巴。



### 标签含义

1:'推广营销'

2:'中文色情赌博'

3:'欺诈钓鱼'

4:' 灰色产业'

5:'英文约会交友'

6:'高风险未判定'

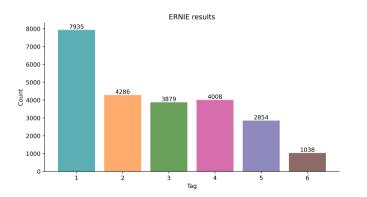
图: Qwen 分类结果



# 分类结果: ERNIE

1 垃圾邮件内容分类实验

ERNIE 开发者: 百度。



# 标签含义

1:'推广营销'

2:'中文色情赌博'

3:' 欺诈钓鱼'

4:' 灰色产业'

5:'英文约会交友'

6:'高风险未判定'

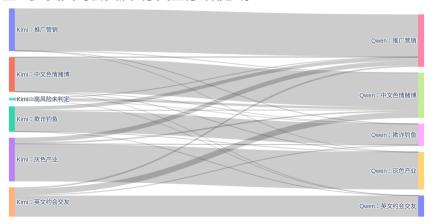
图: ERNIE 分类结果



### 分类结果交叉验证

1 垃圾邮件内容分类实验

"灰色产业"与"英文约会交友"分类上分歧较大。





# 成本分析

1 垃圾邮件内容分类实验

真金白银,请看 VCR:

1. Kimi: 515 元

#### 账户总览

余额 (元): 8.45 元

赠送金额(元):15.00元

昨日消费 (元): 0.00 元

去充值

累计充值(元):515.00元

总消费(元): 521.54 元

本月消费 (元): 0.00 元

图: Kimi API 控制台



### 成本分析

1 垃圾邮件内容分类实验

真金白银,请看 VCR:

1. Kimi: 515 元

2. Qwen: 17元 (限免赠送 4M tokens)

#### 账户总览

余额(元): 8.45元

赠送金额(元): 15.00元

昨日消费 (元): 0.00 元

去充值

累计充值(元):515.00元

总消费 (元): 521.54 元

本月消费 (元): 0.00 元

图: Kimi API 控制台



# 成本分析

#### 1 垃圾邮件内容分类实验

#### 真金白银,请看 VCR:

1. Kimi: 515 元

2. Qwen: 17 元 (限免赠送 4M tokens)

3. ERNIE: 免费

#### 账户总览

余额 (元): 8.45 元

赠送金额(元): 15.00元

昨日消费(元): 0.00元

去充值

累计充值(元):515.00元

总消费 (元): 521.54 元

本月消费 (元): 0.00 元

图: Kimi API 控制台



#### 目录

1 垃圾邮件内容分类实验

▶ 垃圾邮件内容分类实验

基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与利研机构测试



# 实验步骤

1 垃圾邮件内容分类实验

#### 通过 LM Embedding 分类的方法主要分为如下五个步骤:

- 1. 选择特征文本:邮件主题(subject 字段)或邮件正文(content 字段)。
- 2. 选择预训练模型: bert-base-uncased,roberta-base,distilbert-base-uncased,xlnet-base-cased,albert-base-v2 五种预训练模型。
- 3. 文本嵌入向量: 通过预训练模型的 tokenizer 得到特征文本在预训练模型上的 Embedding 向量。
- 4. 降维: 使用 PCA 与 TSNE 两种降维方法观察特征文本相似性。
- 5. 无果:在二维下没有观察出能够有效分类的结果(但是出现了很有意思的形状)。



# 数据组数

1 垃圾邮件内容分类实验

 $2\times5\times2=20$ 

(主题,正文)×(5个预训练模型)×(2种降维)



# 降维结果 demo1

#### 1 垃圾邮件内容分类实验









content\_embedding\_albe rt-base-v2\_(PCA).pdf

content\_embedding\_albe rt-base-v2\_(TSNE).pdf

content\_embedding\_bert -base-unca...d\_(PCA).pdf

content\_embedding\_bert -base-unca...\_(TSNE).pdf









content\_embedding\_xlne t-base-cased\_(PCA).pdf

content\_embedding\_xlne t-base-cased\_(TSNE).pdf

subject\_embedding\_albe rt-base-v2\_(PCA).pdf

subject\_embedding\_albe rt-base-v2\_(TSNE).pdf

图: Embedding 降维结果 demo1



# 降维结果 demo2

#### 1 垃圾邮件内容分类实验









content\_embedding\_disti lbert-base-...d\_(PCA).pdf

content\_embedding\_disti lbert-base-...\_(TSNE).pdf

content\_embedding\_rob erta-base\_(PCA).pdf

content\_embedding\_rob erta-base\_(TSNE).pdf









subject\_embedding\_bert -base-unca...d\_(PCA).pdf

subject\_embedding\_bert -base-unca...\_(TSNE).pdf

subject\_embedding\_disti lbert-base-...d\_(PCA).pdf

subject\_embedding\_disti lbert-base-...\_(TSNE).pdf

图: Embedding 降维结果 demo2



#### 目录

1 垃圾邮件内容分类实验

▶ 垃圾邮件内容分类实验

基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



### 个例研究: 未知错误1

1 垃圾邮件内容分类实验

• ID: pLMsrOCD2F

• 标题: "

• 正文: \r\nwangyiguang1980@bjmu.edu.cnk\r\n\r\nUNSUBSCRIBE EMAIL\r\n\r\n 二零一八年十月十一日下午,美国国会"国际宗教自由核心小组"和非政府机构"国际宗教自由圆桌会议"在美国国会山举行简报会,聚焦中国的宗教信仰自由现况。法轮功发言人张而平说,中共迫害法轮功学员的恐怖罪行仍在持续。他呼吁美国国会和行政部门继续调查在中国仍在发生的强摘器官暴行,并采取一切必要手段制止这一反人类罪行。\r\n 二零一八年七月二十日是中共有系统的暴力打压法轮功学员整整十九年的日子。自一九九九年以来,中共动用整部国家机器迫害在中国修炼法轮功的大约七千万至一亿的



# 个例研究:未知错误2

1 垃圾邮件内容分类实验

• ID: NDVTfGokE1

• 标题: 劝君不必遮人目, 上有苍苍日鉴临

• 正文: \r\n 大哥/大姐: 万事如意! \r\n\r\n\r\n 请看附 | 件,无需在意正文中随机东西,因为主题打不出来,你懂的。\r\n\r\n 后飞班师回朝的消息在百姓之间传开来,百姓哀号痛哭,恳求岳飞不要离去,据说沿路都是跪着哭送岳飞的军民。岳飞回临安后,立即被解除兵权。绍兴十一年(1142 年)秦桧诬岳飞谋反,将其下狱。同年十二月二十九日,秦桧以"莫须有"的罪名,将岳飞毒死于临安风 波亭,岳飞当时年仅三十九岁。\r\n\r\n 像岳飞这样一个忠心爱国的英雄,最后竟被害冤死,且罪名还是荒诞的"莫须有"三个字!岳飞生在这样无奈的年代,配上昏庸的君主、可耻的小人,叫他如何施展



### 个例研究: 高风险未判定 1

1 垃圾邮件内容分类实验

• ID: Udl4oh5nJB

• 标题: 特朗普公布太空導彈防御战 4 略

• 正文: 同学,能联系到您很高兴: \r\n\r\n 英媒 BBC 在 2 月 1 日刊发评论文章说,中 共在新的一年里,对重大风险的警觉性骤然升高。在目前外患内忧的严重困境下,无论是从中 共公 安部 长赵克志召开会议将「颜色革命」作为主要挑战,还是习近平直言中 共要防范化解来自各个方面和领域的重大风险,都表明他们直接感受到了危机或许就在眼前。\r\n\r\n 请读附件,而不要在意正文中的随机内容。\r\n\r\n2018 年中美贸易战爆发以来,中共内部挑战习近平权威的声音不断涌现。在 2018 年夏季的北戴河会议召开前夕,中 共各派不断向外放风,其中一种声音就是要求北京当局应对当前



### 个例研究: 高风险未判定 2

1 垃圾邮件内容分类实验

• ID: HIFyNU85jl

• 标题: 细雨和风陣陣雷, 明媚新天地

• 正文: \r\n 好人,送您吉言: \r\n\r\n\r\n 请读附件,而不要在意正文中的随机内容。\r\n\r\n\r\n 黄帝垂衣裳而天下治\r\n 黄帝是五帝之首,黄帝时期是中国近五千年历史的开端,也是华夏文明的奠基时期。中国人常称为"炎黄子孙",意为炎帝与黄帝是我们华夏民族的共同始祖。\r\n\r\n 以道治国\r\n 黄帝早期以德治国,教化百姓,天下被治理得安宁和谐,百姓幸福。但黄帝并不满足,觉得治理得还不够好,应该可以将天下治理得更好。一次黄帝梦中受到启悟,醒来后悟得治国养身之道,他又四处寻师问道,刻苦修行,不断提升自身境界,最终实现了天下无为而治的最高治世境界。\r\n\r\n 另外舜还弹奏五



2 发件人身份伪造实验

▶ 垃圾邮件内容分类实验 实验总览 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



2 发件人身份伪造实验

▶ 垃圾邮件内容分类实验

买验忌宽 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



#### 全款拿下服务器与域名

- 2 发件人身份伪造实验
- 1. 服务器: 腾讯云免费试用 1 个月https://cloud.tencent.com/act/pro/free
- 2. 域名: 腾讯云购买域名ucashomework.icu
- 3. SSH 连接:配置 SSH 公钥免密登录,好处是可以使用 vscode 了。

```
[16:26:13] [~] >>> ssh gg@ucgshomework.icu
(Welcome to Ubuntu 22.04 LTS (GNU/Linux 5.15.0-106-generic x86_64)
 * Documentation: https://help.ubuntu.com
 * Management:
                  https://landscape_caponical_com
 * Support:
                  https://ubuntu.com/advantage
  System information as of Tue Jun 11 04:26:14 PM CST 2024
  System load: 0.0068359375
  Usage of /: 9.8% of 39.26GB Users logged in:
  Memory usage: 15%
                                 IPv4 address for eth0: 10.2.24.16
  Swap usage:
 * Strictly confined Kubernetes makes edge and IoT secure, Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.
   https://ubuntu.com/engage/secure-kubernetes-at-the-edge
 Last login: Tue Jun 11 16:25:30 2024
 ag@VM-24-16-ubuntu:~$ pwd
/home/aa
Linux VM-24-16-ubuntu 5.15.0-106-generic #116-Ubuntu SMP Wed Apr 17 09:17:56 UTC 2024 x86 64 x86 64 x86 64 GNU/Linux
  a@VM-24-16-ubuntu:~$
```

图: 本机终端 SSH 连接腾讯云服务器



2 发件人身份伪造实验

▶ 垃圾邮件内容分类实验

实验总宽 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果



# Apple 邮箱客户端:成功进入收件箱

2 发件人身份伪造实验

19 类攻击方法中,成功进入收件箱的有如下 7 种。发件人名称显示异常的有如下红框标出的三种。

sumingkang2l@mails.ucas.ac.cn 2024/6/12 A19: Display Name and real address parsing inco Hi, this is a test message! Best wishes.	first@ucashomework.icu A12: From and Sender header ambiguity Hi, this is a test message! Best wishes.
<pre>&lt;@ucashomework.icu &lt;@ucas 2024/6/12 A16: Route portion Hi, this is a test message! Best wishes.</pre>	sumingkang2l@mails.ucas.ac.cn 2024/6/12 A9: Multiple From headers with preceding space Hi, this is a test message! Best wishes.
first@ucashomework.icu 2024/6/12 A13: From and Resent-From header ambiguity Hi, this is a test message! Best wishes.	sumingkang2l@mails.ucas.ac.cn 2024/6/12 A4: DKIM authentication results injection usin  Hi, this is a test message! Best wishes.
	sumingkang2l@mails.ucas.ac.cn 2024/6/12 A1: Non-existent subdomain Hi, this is a test messagel Best wishes.



### Apple 邮箱客户端: 个例研究

2 发件人身份伪造实验



#### sumingkang2l@mails.ucas.ac.cn

A1: Non-existent subdomain 收件人: liuxun21@mails.ucas.ac.cn

Hi, this is a test message! Best wishes.



#### sumingkang2l@mails.ucas.ac.cn <any@ucashomework.icu>

A19: Display Name and real address parsing inconsistencies 收件人: liuxun21@mails.ucas.ac.cn

Hi, this is a test message! Best wishes.

图: A1 攻击结果

图: A19 攻击结果

- A1 攻击完全成功,也没有显示"代发"。
- A19 是唯一一个头像显示首字母的类型。但是昵称后会显示攻击者域名。



2 发件人身份伪造实验

▶ 垃圾邮件内容分类实验

实验总觉 基于 LLM Chat 的垃圾邮件分类 基于 LM Embedding 的垃圾邮件分类 个例研究

► 发件人身份伪造实验 前期准备:全款拿下 Apple 邮箱测试结果 跨高校与科研机构测试



# 多机构测试: ICT 计算所

2 发件人身份伪造实验

• 结果: 失败。服务器直接拒绝。

>>> 354 End data with <cr><lf>.<cr><lf></lf></cr></lf></cr>
<<< From: <admin_ict@ofey.com></admin_ict@ofey.com>
To: <sunfei@ict.ac.cn></sunfei@ict.ac.cn>
Subject: A1: Non-existent subdomain
Date: Tue, 11 Jun 2024 04:53:07 +0000
Content-Type: text/plain; charset="UTF-8"
MIME-Version: 1.0
Message-ID: <1538085644648.096e3d4e-bc38-4027-b57e-TQW6XW@message-ids.http://ucashomework.icu
X-Email-Client: https://github.com/chenjj/espoofer
Hi, this is a test message! Best wishes.
>>> 550 Requested mail action not taken: Invalid IP<82.156.151.99>
<<< quit
>>> 221 Bye
Finished.

图: 计算所



#### 多机构测试: UCB 加州大学伯克利分校

2 发件人身份伪造实验

• 结果: 成功。



图: 伯克利



## 多机构测试: THU 清华大学

2 发件人身份伪造实验

• 结果: 成功。

#### **Admission to Stanford University**

official@stanford.cool 发送给 msh@233.com

₽ ⑤ 號 Ё 发起会议 2024-06-11 17:13:16



AQ is thrilled to inform you that you have been admitted to the Class of 2028 at Stanford University! This is a remarkable accomplishment, and we commend you on your exceptional achievements and hard work.

图: 清华大学



### 多机构测试: PKU 北京大学

2 发件人身份伪造实验

结果:成功。



#### official@stanford.cool 收件人: fv@233.com >

19:12

#### **Admission to Stanford University**

AQ is thrilled to inform you that you have been admitted to the Class of 2028 at Stanford University! This is a remarkable accomplishment, and we commend you on your exceptional achievements and hard work.

图: 北京大学



### 多机构测试: ZJU 浙江大学

2 发件人身份伪造实验

• 结果: 半成功。进入垃圾箱。

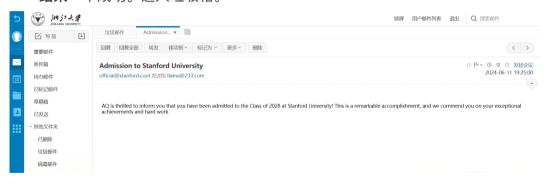


图: 浙江大学



#### 多机构测试: FDU 复旦大学

2 发件人身份伪造实验

• 结果: 半成功。邮箱客户端有提示。



图: 复旦大学



# 垃圾邮件一条龙:从伪造到分类

大作业完结撒花!

任何问题?