25 Tweets to Know You: A New Model to Predict Personality with Social Media

Research Objective

人们在社交网络上的行为可以反映人格特征，同时人格特征也会影响人们在社交网络的行为。

本文研究目标就是**如何利用社交网络上的信息预测用户的人格特征.**

Problem Statement

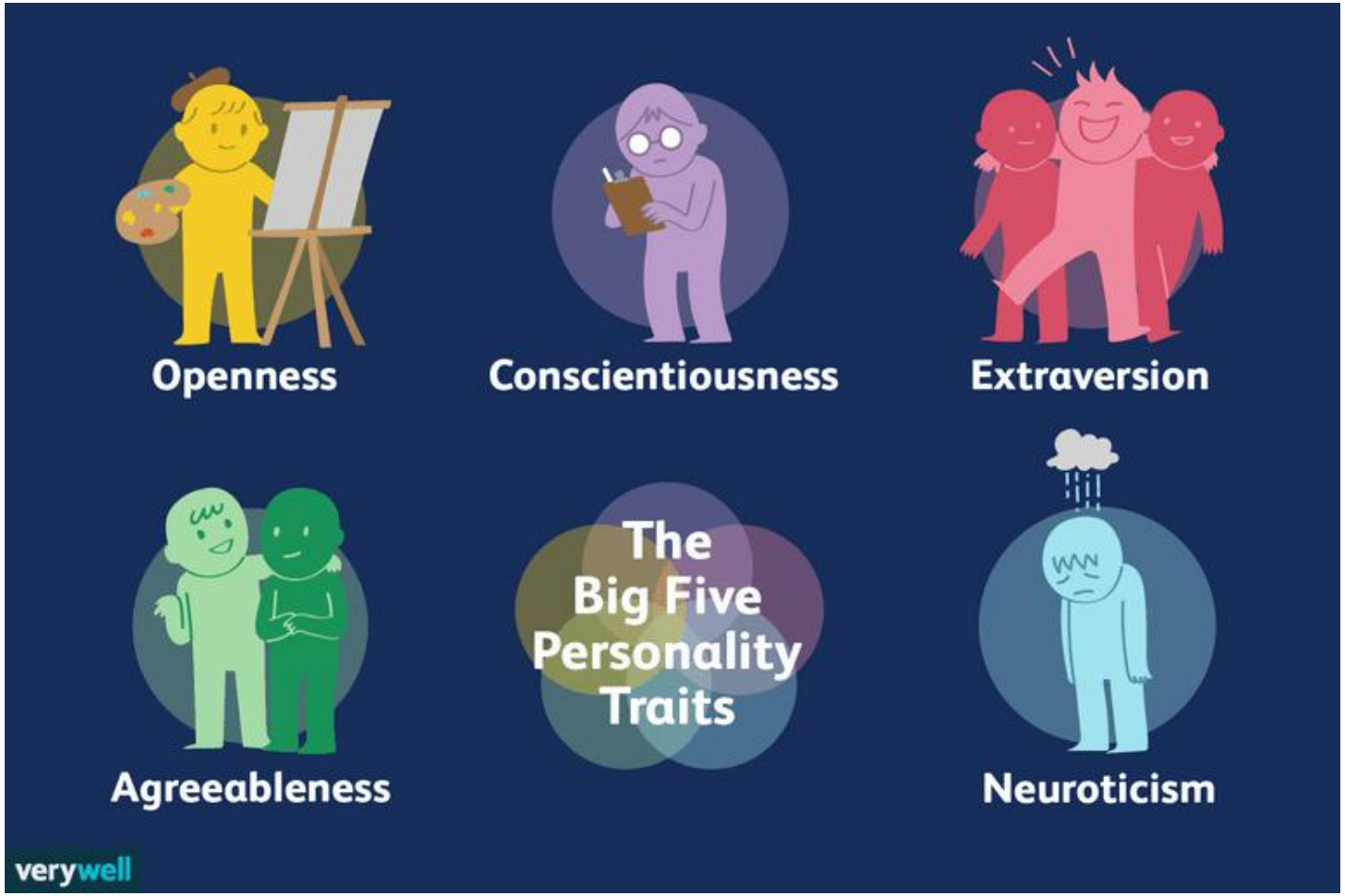
之前的方法由于需要太多的数据，而导致不能可靠的预测用户人格特征。因此，引出如何利用少量的社交媒体数据来预测。

Priori Knowledge

AMC
AI ML Club

大五人格

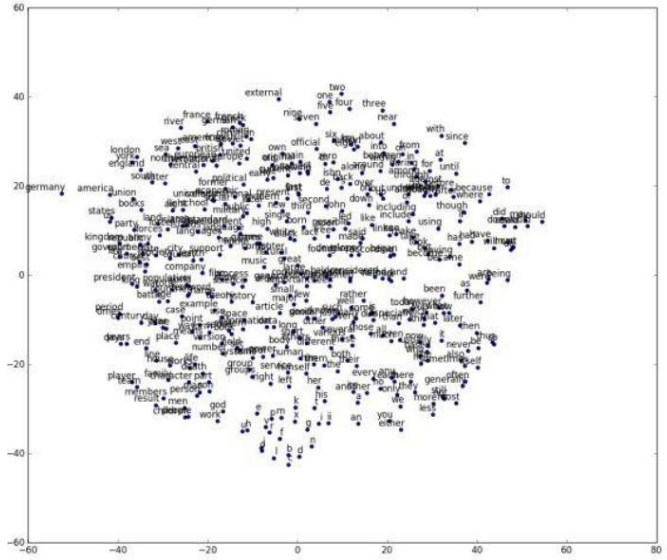
- **开放性（openness）：** 具有想象、审美、情感丰富、求异、创造、智能等特质。
- **责任心（conscientiousness）：** 显示胜任、公正、条理、尽职、成就、自律、谨慎、克制等特点。
- **外倾性（extroversion）：** 表现出热情、社交、果断、活跃、冒险、乐观等特质。
- **宜人性（agreeableness）：** 具有信任、利他、直率、依从、谦虚、移情等特质。
- **神经质性（neuroticism）：** 难以平衡焦虑、敌对、压抑、自我意识、冲动、脆弱等情绪的特质，即不具有保持情绪稳定的能力



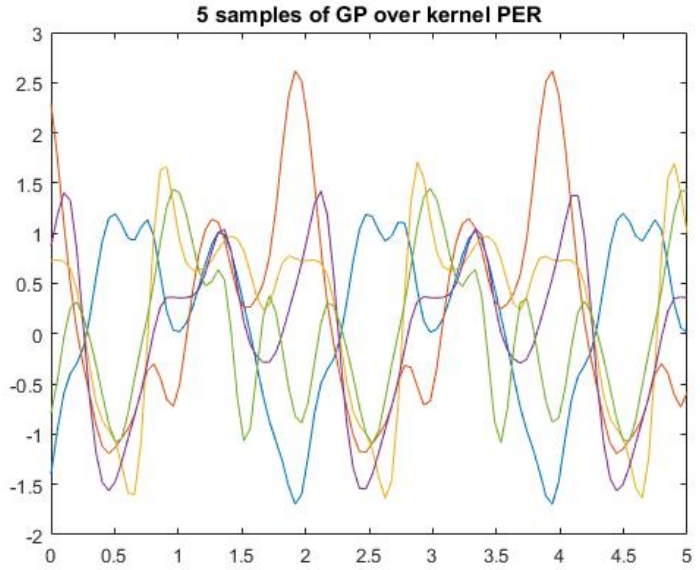
Priori Knowledge

AMC
AI ML Club

Word Embedding: 将单词映射至某个语义空间中$f: X \rightarrow Y$



Gaussian Process:一系列关于连续域（时间或空间）的随机变量的联合，而且针对每一个时间或是空间点上的随机变量都是服从高斯分布的

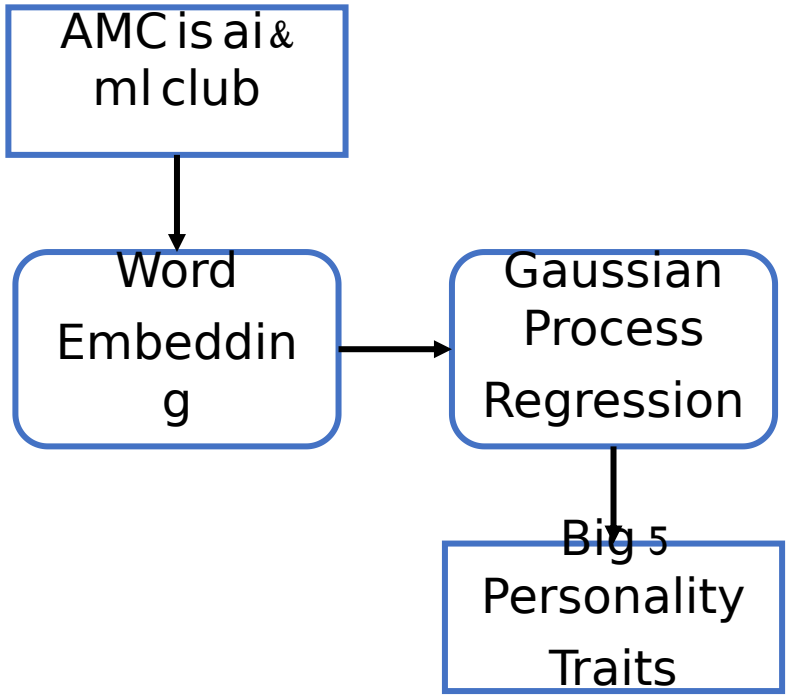


Methods

提出集成Word Embedding特征的高斯过程回归模型。

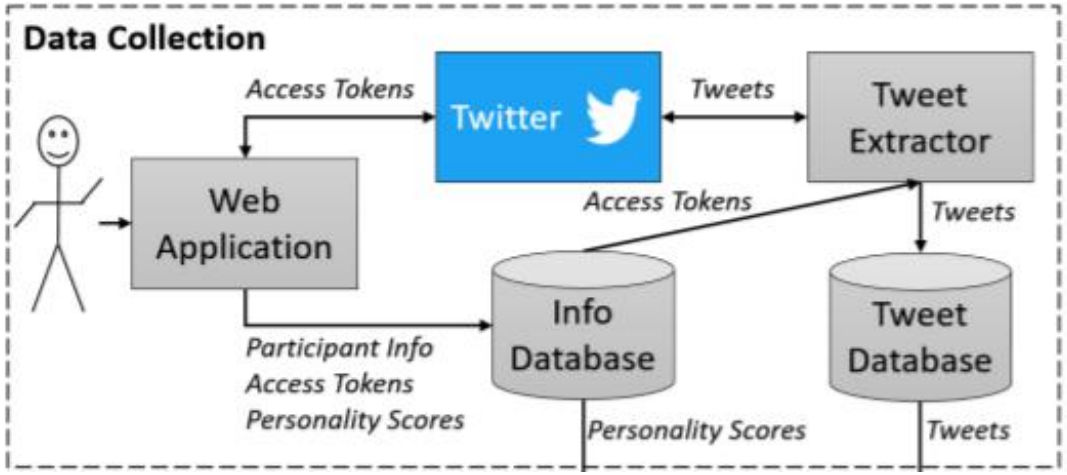
具体步骤：

1. 将抽取用户推文的单词，平均所有每个词的word embedding到单个向量上(200 dim)
2. 高斯过程模型接收该向量作为模型输入，输出Big-5的预测结果



Experiments

数据收集
由于人格特征是针对用户而言，所有需要对用户进行标注。 作者设计了web app以广告的形式投放给推特用户。通过web app，让用户自愿同意分享他们的推文并回答一项性格调查。



总共收集了1323名用户
用户年龄分布： 18 to 24 (47%), 25 to 34 (14%), 35 to 54 (12%), above 54(3%)
Big5得分分布： O[开放性]($\mu = 0.76, \delta$ =0.12), C[责任心]($\mu = 0.59, \delta = 0.15$), E[外倾性]($\mu = 0.54, \delta = 0.18$), A[宜人性]($\mu$ =0.72, δ = 0.13), N[神经质](μ = 0.44, δ = 0.19).

Experiments

对比的方法：

- Linguistic Inquiry and Word Count(LIWC)+ Rigde Regression(RR)
- 3-Gram + Rigde Regression(RR)
- Word Embedding(GloVe) + Gaussian Process(GP)

Metric:
度量预测人格特征得分与ground-truth的皮尔逊相关系数

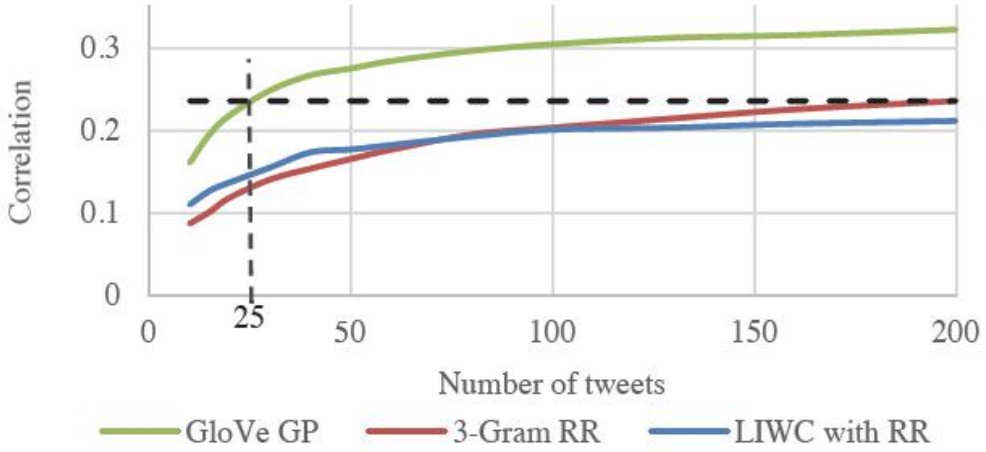


Figure 1. Prediction accuracy of the Big-5 traits according to the number of tweets. Reported correlations are significant $p < 0.01$.

	Agree.	*Consc.*	*Extrav.*	*Neurot.*	*Openn.*
LIWC RR	0.25	0.27	0.13	0.28	0.26
3-Grm RR	0.21	0.28	0.18	0.35	0.26
GloVe RR	0.27	0.27	0.27	0.27	0.27
LIWC GP	0.26	0.28	0.20	0.33	0.26
3-Grm GP	0.11	0.12	0.15	0.15	0.09
GloVe GP	0.29	0.33	0.25	0.42	0.37

Table 1. Model correlation comparison for the Big-5 traits. The reported correlations are significant $p<0.01$.

Conclusion

AMC
AI ML Club

- 作者提出的方法（Word Embedding +Gaussian Process Regression）在所有三个实验中都优于目前最先进的人格预测方法，并且可以使用更少的数据达到较好的效果
- Word Embedding 特征非常适合高斯过程（GP）：由于其内部的内核表示，GP依赖于特征的协方差。

Thanks!