工具：LIWC（语言查询和字数统计）

基于EmoGraph图的方法

MapReduce

Estival

N-Gram模型：N-Gram是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为N的滑动窗口操作，形成了长度是N的字节片段序列。

通过用户关联分析 消息中包含@符号，@后跟了人名id，可以通过对id的提取判断性别后关联性别

二分类问题

n-元词作为特征 别准确率(71.2%)

n-元词特征与社会语言风格特征（语言风格特征主要包括表情符号，省略号，语气词等） 准确率为 72.3%

tf-idf加权n-gram特征

文件断词算法：(N-gram Segmentation Algorithm, NSA)

N-gram 的文件斷詞演算法(N-gram Segmentation Algorithm, NSA)

使用機率式詞頻與反類別頻率 (Probability based Term Frequency and Inverse Category Frequency, PTFICF)將特徵值加 權，最後使用 SVM 訓練萃取出特徵值

詞頻反轉文件頻系數(Term Frequency – Inverse Document Frequency, TFIDF)，

预处理：停用词过滤、还原字根、N-Gram元词、特权值转换

