# 汽车行业用户观点主题及情感识别

## ——数据科学导论期末报告

**何秦兴** PB16110299

**吴雨菲** PB15020525

**张劲暾** PB16111485

2018 **年** 12 **月** 6 **日**

# 目录

# 1 题目描述

## 1.1 题目背景

随着政府对新能源汽车的大力扶植以及智能联网汽车兴起都预示着未来几年汽车行业的多元化发展及转变。汽车厂商需要了解自身产品是否能够满足消费者的需求，但传统的调研手段因为样本量小、效率低等缺陷已经无法满足当前快速发展的市场环境。因此，汽车厂商需要一种快速、准确的方式来了解消费者需求。

## 1.2 题目任务

本赛题提供一部分网络中公开的用户对汽车的相关内容文本数据作为训练集，训练集数据已由人工进行分类并进行标记，参赛队伍需要对文本内容中的讨论主题和情感信息来分析评论用户对所讨论主题的偏好。讨论主题可以从文本中匹配，也可能需要根据上下文提炼。

## 1.3 数据说明

训练集数据中主题被分为 10 类，包括：动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性。

情感分为 3 类，分别用数字 0、1、-1 表示中立、正向、负向。

content_id 与 content 一一对应，但同一条 content 中可能会包含多个主题，此时出现多条记录标注不同的主题及情感，因此在整个训练集中 content_id 存在重复值。

| 字段名称 | 类型 | 描述 | 说明 |
|---|---|---|---|
| content_id | Int | 数据 ID | / |
| content | String | 文本内容 | / |
| subject | String | 主题 | 提取或依据上下文归纳出来的主题 |
| sentiment_value | Int | 情感分析 | 分析出的情感 |
| sentiment_word | String | 情感词 | 情感词 |

## 1.4 评测标准

本赛题采用 F1-Score 评价方式。按照“主题 + 情感分析”识别数量和结果（是否正确）来进行判断，参赛者需要识别文本中可能包含的多个“主题”。匹配识别结果: Tp: 判断正确的数量；Fp: 判断错误或多判的数量；Fn；漏判的数量.

当提交的一条数据结果包含“主题 + 情感值”，如果参赛者对“主题 + 情感”的判断结果完全正确则计入 Tp，如果对“主题”或“情感值”的判断结果错误则计入 Fp；如果参赛者未能对某一数据文本判断“主题”或“情感值”给出判断结果，则此条数据不能包含在结果文件中；如果参赛者识别出的“主题 + 情感值”数量少于测试样本中实际包含的数量，或未对某个测试样本数据给出结果，缺少的数量计入 Fn；如果参赛者识别出的“主题 + 情感值”数量多于测试样本中实际包含的数量，超出的数量计入 Fp

$$\text{准确率：} P = \frac{T_p}{T_p + F_p}$$

$$\text{召回率：} R = \frac{T_p}{T_p + F_n}$$

$$\textbf{F1-Score：} F1 = \frac{2PR}{P + R}$$

# 2 题目分析与方案设计

## 2.1 对数据的基本特征分析和设计思路

数据基本特征和任务分析：短文本情感分类与主题挖掘，通过将文本映射到向量空间，并将需要预测的情感和主题用独热码表示或编号为类别，将问题转化为学习从文本特征表达到类标签的的映射关系。

文本的数量特征对模型设计的影响：文本长度普遍较短，可以提取的特征较少，所以需要提取有效的特征；多标签问题：每个实例可能对应多个主题标签，而且对于不同的主题标签还有可能是不同的情感分类，这样多主题的数据大概占到总数据的 15%，这一点对于我们采用的主题、情感分开学习，不区分同一 ID 的不同数据实例的设计方案是不太友好的，参考成功的比赛选手的方案，应该将情感和主题组合得到 30 个标签进行学习，以解决多标签问题。数据数量偏少，只有不到一万条；数据质量差，很多主题情感标签模糊，对于很多正面或者负面的评论笼统标注为中性，中性情感标签明显多余另外两类，造成一定的学习过拟合问题，特别是对于 LSTM 模型的学习影响较大。

## 2.2 特征提取与模型选择

对与文本的特征映射，我们采取了两个大的思路方向，第一是采取将文本分词之后映射到一维的特征向量 LDA

# 3 方法说明

## 3.1 分词

我们首先去除了非中文符号，并且找到了一份通用的汉语停用词表，人工去除了其中

### 3.2 特征提取

#### 3.2.1 隐含狄利克雷主题模型 (LDA)

#### 3.2.2 词袋模型（BOW）

#### 3.2.3 TFIDF

#### 3.2.4 输入 LSTM 的词向量矩阵

### 3.3 学习方法

#### 3.3.1 逻辑回归模型（LR）

#### 3.3.2 朴素贝叶斯模型（NB）

#### 3.3.3 多层感知机模型（MLP）

#### 3.3.4 随机森林模型（RF）

#### 3.3.5 集成模型（Ensemble）

#### 3.3.6 长短期循环神经网络（LSTM）

### 3.4 提升和改进的思路

## 4 试验结果

### 4.1 线上比赛结果

### 4.2 各种方法的结果与分析

### 4.3 没有解决的问题和主要困难

## 5 附录

### 5.1 个人收获

吴雨菲:

何秦兴:

张劲暾:

### 5.2 分工说明

吴雨菲:

何秦兴:

张劲暾:

## 6 参考文献