**基于主题的微博情感分析**

作 者：黎国本 刘振兴 吴俊标 麦驿峰 李炎浩

班 级：软件工程1503指导老师：蒋盛益报告提交时间：2017年12月8日

目录

[摘要 3](#_Toc500621061)

[一、 绪论 3](#_Toc500621062)

[1.1 研究背景 3](#_Toc500621063)

[1.2 研究意义 4](#_Toc500621064)

[1.3 可行性分析 4](#_Toc500621065)

[（1） 数据真实 4](#_Toc500621066)

[（2） 选取的指标具有较高的科学性及可操作性 4](#_Toc500621067)

[（3） 情感分析的准确率 4](#_Toc500621068)

[1.4 实验过程 4](#_Toc500621069)

[（1） 收集数据 4](#_Toc500621070)

[（2） 数据预处理 4](#_Toc500621071)

[（3） 发展模型与假设 5](#_Toc500621072)

[（4） 解释和使用数据以及对数据挖掘结果的评估 5](#_Toc500621073)

[1.5 本文的创新之处 5](#_Toc500621074)

[二、 运用xx算法对某话题下的微博进行情感分析 5](#_Toc500621075)

[2.1 研究对象 5](#_Toc500621076)

[2.2 样本选择 6](#_Toc500621077)

[（1） 本课题选定的话题 6](#_Toc500621078)

[（2） 数据来源 6](#_Toc500621079)

[（3） 数据属性 6](#_Toc500621080)

[（4） 样本示例 6](#_Toc500621081)

[2.3 指标选择 6](#_Toc500621082)

[2.4 算法流程 7](#_Toc500621083)

[（1） 相关依赖 7](#_Toc500621084)

[（2） 具体流程 8](#_Toc500621085)

[2.5 可视化 9](#_Toc500621086)

[2.6 结果分析 9](#_Toc500621087)

[2.7 实验结论 13](#_Toc500621088)

[三、 本课题局限性 16](#_Toc500621089)

[四、 总结与体会 16](#_Toc500621090)

[五、 成员分工与自评 16](#_Toc500621091)

[5.1 成员分工 16](#_Toc500621092)

[5.2 自评 17](#_Toc500621093)

[六、 参考文献 17](#_Toc500621094)

# 摘要

随着微博的风靡，与之相关的研究得到学术界和工商界的广泛关注。针对中文微博情感分析的研究进行综述。将中文微博文本情感分析分为四个步骤：数据收集、数据预处理、情感信息抽取和情感分类，对各自的研究方法和进展进行总结。其中微博文本情感分类方法主要采用了word2vec与机器学习模型的组合使用的情感分类，分类器则是采用了支撑向量机和朴素贝叶斯。最后将分析结果可视化，直观明了。

**Abstract**

With the popularity of micro-blog, the related research has received extensive attention from the academia and the business community. The research on Chinese micro-blog emotion analysis is reviewed. The Chinese micro-blog text sentiment analysis is divided into four tasks: data collection, data preprocessing, emotion information extraction and emotion classification, and their respective research methods and progress are summarized. Among them, micro-blog text sentiment classification method adopts the combination of word2vec and machine learning, and the classifier adopts support vector machine and naive Bayes. Finally, the results of the analysis are visualized and it looks clear and clear.

**关键词**：情感分类 词袋 支撑向量机 Tf-idf hashmap Word2vec 朴素贝叶斯

# 绪论

## 研究背景

微博，即微型博客的简称，是一个基于用户关系信息分享、传播以及获取的社交网络平台。用户以140字（包括标点符号）的文字更新信息，并实现即时分享。自其问世以来，大量用户使用微博记录生活、讨论热点话题、表达和分享观点。而微博上信息繁多且增长速度很快，仅靠传统人工方法难以应对海量信息的收集和处理工作，无法高效地获得有价值的信息，迫切需要数据挖掘技术对其进行处理和分析。情感分析技术能够自动将文本中表达的情感倾向进行正负面的分类，很大程度上解决微博上信息杂乱的现象，方便用户快速准确定位所需信息。

## 研究意义

所谓情感分析，是指分析说话者在传达信息时所隐含的情绪状态，对说话者的态度、意见进行判断或者评估。情感分析在微博海量数据上的应用，将有助于完善互联网的舆情监控系统；丰富和拓展企业的营销能力；通过波动分析，实现对物理世界异常或突发事件的检测。此外，还可以应用于心理学、社会学、金融预测等领域的研究。且微博时刻传递着海量信息，是一个很好的研究对象。故对于微博情感分析的研究有着很重要的现实意义。

## 可行性分析

### 数据真实

本课题数据来自于新浪微博，使用小组自行编写的网络爬虫获取得到，且数据均为近期的微博话题内容，具有时效性。

### 选取的指标具有较高的科学性及可操作性

选取指标主要从两个方面进行考虑，一是是否有相对应的训练集，二是是否能给我们的分析结论提供分析帮助。在做微博情感分析，有使用单一的正负情感分析，也有多分类的情感。从理论上来说，都是具有可操作性，所以我们决定做这两部分的实验。

### 情感分析的准确率

一般来说，正负情感分析准确率较高。而多分类的情感分析准确率较低。

## 实验过程

### 收集数据

本课题中，我们编写了基于scrapy框架的微博网络爬虫进行数据的收集工作，爬取了大量数据，即选定的（一些近期热门的微博话题）微博内容、评论内容、相关用户信息等等，并对数据清洗后将其存入MySQL数据库中。

### 数据预处理

我们选择使用分类的方法，所以我们先获取了两个数据集用于来训练我们的模型。一个

是只有标记正负情感的训练集，约一万条，正负情感各5000条。一个是NLPCC2014的微博情感分析的数据集，一共分为七类情感，数据不平衡，共40000多条。

1. xml转为txt
2. 分离文本和分类标签
3. 对文本去除停用词
4. 分析选择分词方法
5. 转化成矩阵
6. 词袋（稀疏矩阵）
7. Tf-idf
8. 特征哈希
9. Word2vec

### 发展模型与假设

对于这类分类问题，通常会选择传统的分类器。

1. 线性支撑向量机
2. 朴素贝叶斯
3. K折交叉验证

### 解释和使用数据以及对数据挖掘结果的评估

1. 评测模型的方法
2. 精确率precision
3. 召回率recall
4. F1值
5. 可视化方式
6. 图表
7. 词云

## 本文的创新之处

本文的实际工作其实更倾向于对微博这种类型的短文本进行情感分类。尝试了许多不同的方法。

创新之处：

* 1. 尝试了word2vec与机器学习模型的组合使用
  2. 尝试的做了多步分类，先进行正负情感分类，再做多分类。

# 对特定话题的微博进行情感分析

## 研究对象

选定话题下的微博情感状态。

## 样本选择

### 本课题选定的话题

1. 鹿晗关晓彤
2. 运动教室
3. 致我们单纯的小美好
4. 热搜疯狂打call

### 数据来源

新浪微博

### 数据属性

1. 微博ID
2. 微博内容
3. 微博发布时间
4. 发布该微博的用户ID
5. 转发数；
6. 评论数；
7. 点赞数；
8. 评论ID；
9. 评论内容；
10. 该评论指向的微博ID/评论ID；
11. 发布该评论的用户ID；
12. 评论发布时间

### 样本示例

数据样本如图1（以鹿晗关晓彤话题为例）：

C:\Users\zhenxing\AppData\Local\Temp\1512791073(1).png

图1 数据样本示例

注：本课题使用的数据均由小组自行开发的基于scrapy框架的网络爬虫不间断获取。

## 指标选择

主要分为两类指标：

第一类：Positive 和 Negative，如图2所示：

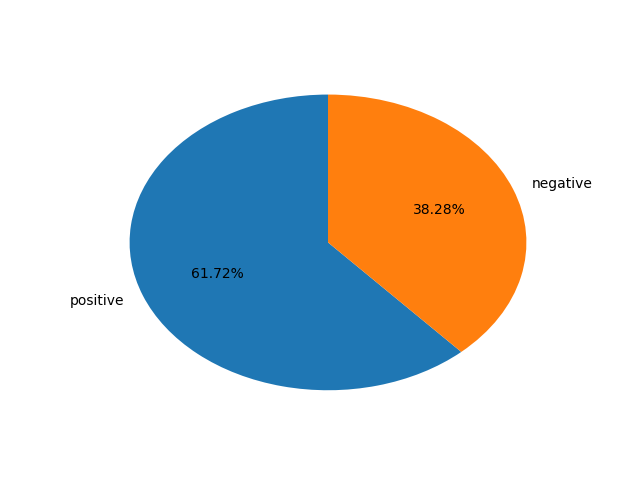


图2 指标示例-第一类

第二类：Null、Like、Anger、Disgust、Happiness、Sadness和Fear，如图3所示：

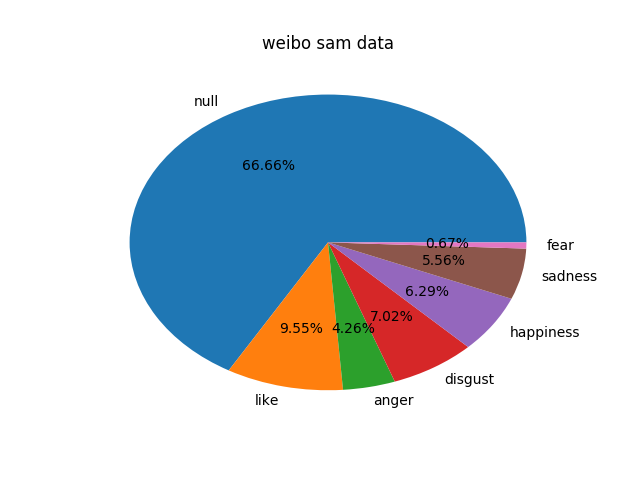


图3 指标示例-第二类

## 算法流程

算法全部通过代码实现，没有使用spss或者clementine软件。

### 相关依赖

* 1. ubuntu16.04 & python2.7
  2. scikit-learn
  3. jieba分词& standford parser & 哈工大ltp
  4. keras
  5. gensim3.1
  6. Numpy & pandas

### 具体流程

* 1. 训练数据获取：

主要有两个训练集，一个是只有正负情感的平衡数据集共10000条。另一个NLPCC2014微博情感评测的数据集，包括（Null、Like、Anger、Disgust、Happiness、Sadness和Fear）七类情感的数据共40000条，其极其不平衡。

* 1. 训练数据集预处理：
  2. 通过python的xmltree解析工具，把xml文件中我们需要的数据分类label和文本数据提取出来，保存到train.txt中。
  3. 使用列表或者字典去读取train.txt，并且把文本和其对应的label分离开。
  4. 在短文本或者是网络上的文本中，会有各种各样奇怪的符号，这些符号会严重干扰后续的特征选择，所以我们选择在这里去除。
  5. 特征工程（特征选择和特征抽取）：
  6. 首先比较了几种分词的工具（结巴分词，standford parser ，哈工大ltp），对于文本基本都是中文的微博来说，我们最后使用了结巴分词作为分词，对训练集进行分词。
  7. 先使用了sklearn中的CountVectorizer方法和TfidfVectorizer去转化词袋，由于词袋有60000\*40000的巨大矩阵，在8g内存的电脑中无法同时读入。
  8. 为了继续进行实验，我们只能换其他的特征抽取方法。首先我们选择放弃不平衡数据当中的null标签的数据，因为其占了大量的数据量，这也是数据不平衡的一大原因。在去除了null标签的数据后，训练集合剩余大约16000条，这样我们可以使用CountVectorizer方法和TfidfVectorize方法去转化为词袋，作为未来放入分类器的特征。
  9. 另一种方式是使用HashingVectorizer进行降维，把降维后的矩阵作为分类器的特征。
  10. 对分词的结果做词性标注，将其转化为bool值，和上述的特征组合或者单独作为特征放入分类器分类。
  11. 改变上述的词袋模型，而改用word2vec，用gensim训练词向量，把词向量作为分类的特征。
  12. 在实验过程中，我们对多情感训练集试用了上述的所有方法，对正负情感只用了词向量。
  13. 对此分类我们主要用了两种分类器（支撑向量机和朴素贝叶斯）：
  14. 主要使用了sklearn进行分类器的构建。
  15. 对于支撑向量机我们尝试了调整惩罚系数c，和内核选择过rbf以及linear。
  16. 同样使用了sklearn进行了10择交叉验证
  17. 对分类器的结果进行了精确率，召回率和f1值的计算。
  18. 通过matplotlib和wordcloud进行了可视化。

## 可视化

为了使分析结果更加直观与具体，我们做了系统的可视化部分，具体的形式是以

1. Web网页的形式呈现。
2. 考虑到数据交互以及模块化开发，网页的主体框架选择了Vue。响应式部分使用了目前较为流行的Bootstrap框架来实现响应式的页面。
3. 呈现数据方面我们选用了E-chart来实现各种数据统计图。

## 结果分析

* 1. 词袋结果（SVM），如图4所示：

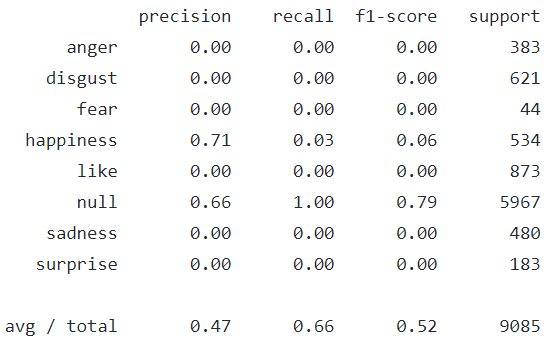


图4 词袋结果（SVM）

* 1. Tf-idf结果（SVM），如图5所示：

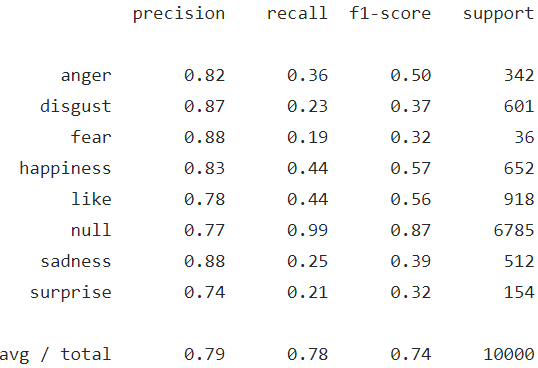


图5 Tf-idf结果（SVM）

* 1. Hashing结果，如图6所示：

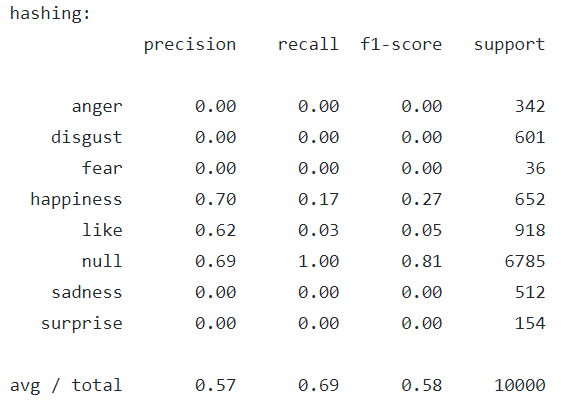


图6 Hashing结果

* 1. 多项式贝叶斯结果，如图7所示：

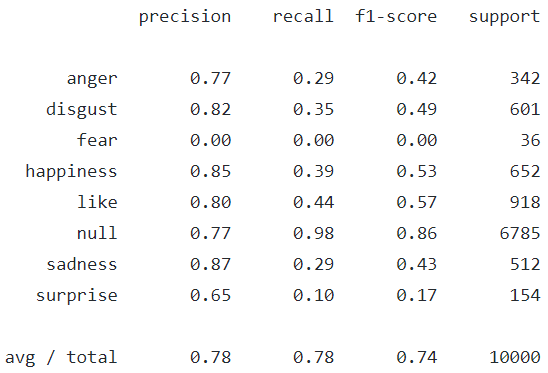


图7 多项式贝叶斯结果

* 1. 伯努利贝叶斯结果，如图8所示：

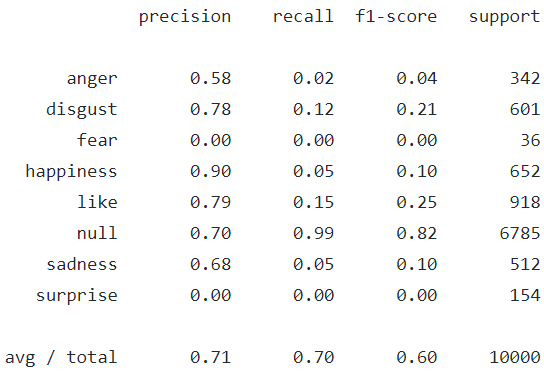


图8 伯努利贝贝叶斯结果

* 1. Word2vec结果（SVM），如图9所示：

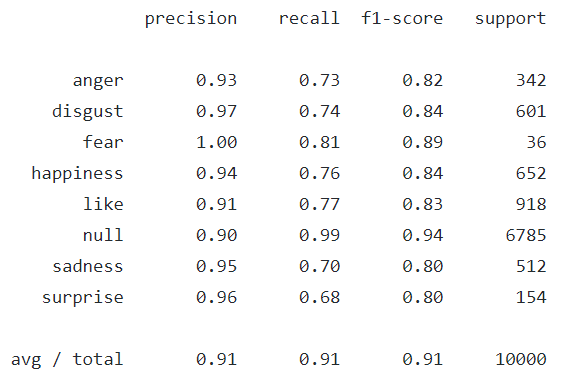


图9 Word2vec结果（SVM）

* 1. Word2Vec结果（SVM 没有null），如图10所示：

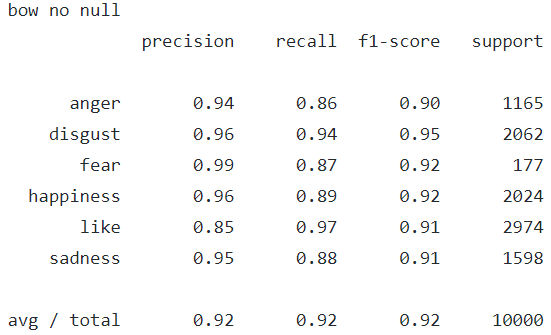


图10 Word2vec结果（SVM 没有null）

分析：

我们可以发现只用词袋作为模型的时候，不管我们做了一些权重的改变工作，或者是使用了LDA和PCA去进行降维，得到的分类效果都不是很好，在微博这种短文本当中似乎很难抽取有效的特征。当我们尝试训练词向量，将其作为特征时，与支撑向量机相结合反而得当了很不错的效果。

当然在实验中也还有很多不尽人意的地方，比如我们的数据极度不平衡。特别是“NULL”，也就是无情感的文本太多，让我们的分类方法很难完善。微博的表情的情感分析也是一个十分难的问题，在我们这次的分类中，只是简单的转化成了文本，没有进行太多的处理。

还有一个不足之处就是没有能够基于不同的主题进行分析，而是针对所有微博的这些短文本做了分类。要想根据不同的主题做出变化，我们可能需要更多不同的特征和方法，比如用主题模型等。

## 实验结论

以#鹿晗关晓彤#话题为例，我们用训练好的模型进行分析和可视化，如下列图所示。

1. 在该话题下的言论的六类情感比例图（图11）：

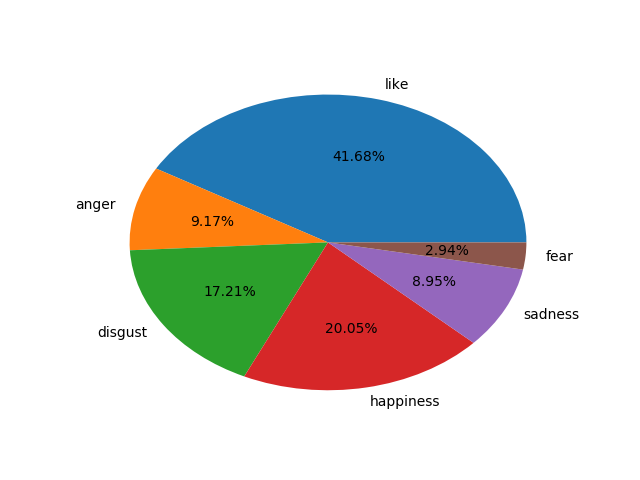


图11六类情感比例图

1. 在该话题下的言论的六类情感数值柱状图（图12）：

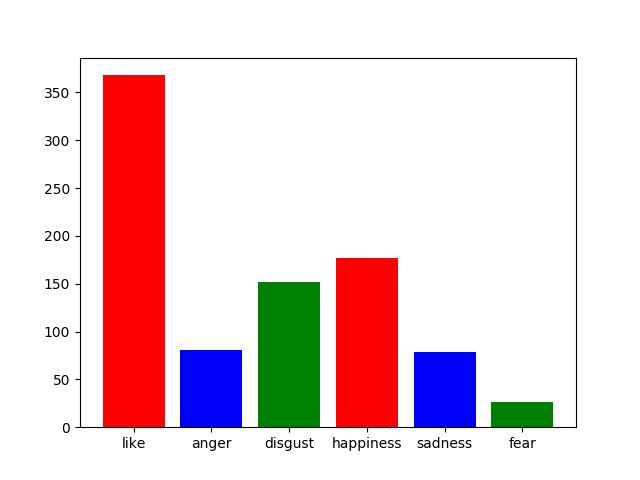


图12六类情感数值柱状图

1. 在该话题下的言论的正负情感比例图（图13）：

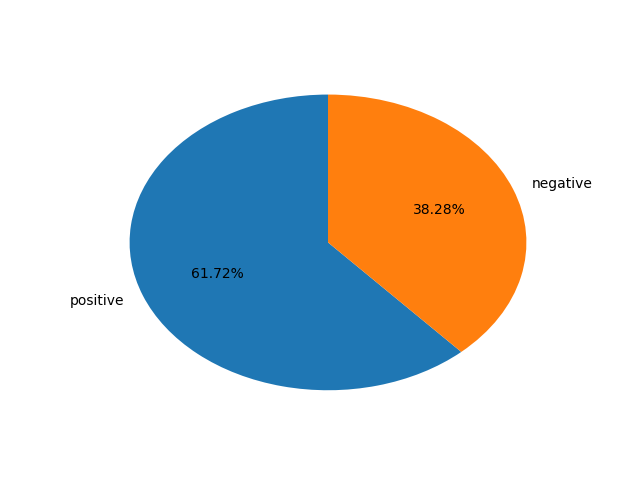


图13正负情感比例图

1. 在该话题下的言论的正负情感数值柱状图（图14）：

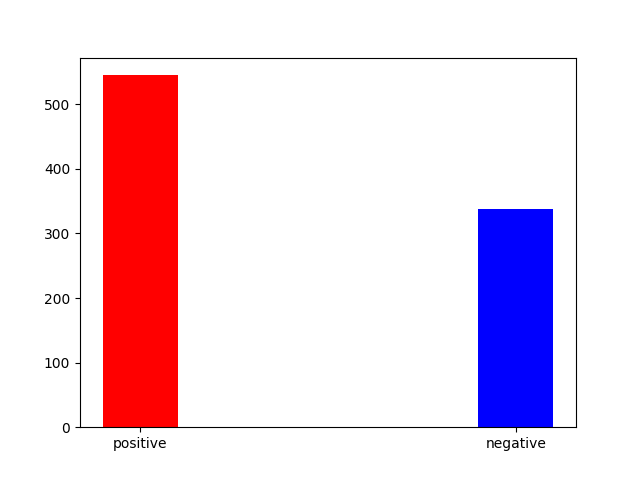


图14正负情感数值柱状图

1. 在该话题下的关键词词云（图15）：



图15关键词词云

结论：

我们可以通过我们的模型和可视化工具代码，对不同的主题生成这一些可视化的图片。以这个话题为例子，从这些图片我们可以看出，对于鹿晗关晓彤这对明星情侣，正负面情绪大概是6：4，可以看出有不少网民对该事件并不看好。在负面情绪中，有不少人感到伤心和反感，也就是说有不少人对该事件表示得十分失望或许不在喜欢他们的偶像。通过词云，可以看到针对该事件的关键词，关键事件和关键情感。

# 本课题局限性

1. 未对水军进行检测，在一些比较大的事件中，特别是负面事件会有很多水军带节奏。由于我们的模型和技术有限，没有办法将水军分辨出来，对结果的分析也会有一定的影响。
2. 未对文本描述的对象进行检测，像微博这种短文本，很难去分辨这些文本的主题和评价的对象问题，所以很难做到真正针对主题和对象的模型构建。
3. 选取的指标不尽合理，无法具体且准备地描述微博情感。人们的情感肯定不止这一些，还有就是无情感的表达，我们都没办法太完善，一定的影响的准确性。

# 总结与体会

总体来说我们用代码实现了全部流程并且得到了一个比较好的结果。在我们用代码实现爬虫、特征选择、模型构建、可视化的过程中遇到了很多的问题。但是都通过我们的团队合作解决了问题。在整个过程中学到了很多相关的知识。

但是我们也有些欠考虑的地方，从一开始想做基于主题的微博情感分析，到最后实际上做得是微博类型的短文本情感分析，以及一些技术上的问题都做得不够完善。

# 成员分工与自评

## 成员分工

课题选取与分析：全体成员

数据格式、指标选取：黎国本、刘振兴、麦驿峰、吴俊标

算法、建模：刘振兴、黎国本

微博爬虫：黎国本、麦驿峰

前端：李炎浩

后台：吴俊标

文档：麦驿峰、刘振兴、吴俊标

## 自评

黎国本：95分 吴俊标：95分 麦驿峰：95分 刘振兴：95分 李炎浩：95分

# 参考文献

[1] scikit-learn. <http://scikit-learn.org>

[2] numpy. <http://www.numpy.org/>

[3] scipy. <https://www.scipy.org/>

[4] Python Data Analysis Library. [http:/pandas.pydata.org/](http://pandas.pydata.org/)

[5] World Cloud Generator. <https://www.jasondavies.com/wordcloud/>

[6] matplotlib. <http://matplotlib.org/>

[7] Chinese-sentiment-analysis. <https://github.com/Zbored/Chinese-sentiment-analysis>

[8] Word Clouds in Python. <https://github.com/dmarklein/WordCloud>

（注：本项目所有代码已上传github：https://github.com/Lauzanhing/WeiBo\_SentimentAnalysis）