中文微博情感分析评测大纲(修订版)

1. 评测对象

本次评测的对象是面向中文微博的情感分析核心技术，包括观点句识别、情感倾向性分析和情感要素抽取。

2. 任务设置

本评测设置了如下 3 个子任务，其中任务 1 是必选任务，任务 2 和任务 3 都是基于任务 1 的，参赛队伍可以选做。

2.1 观点句识别

针对每条微博中的各个句子，本任务要求判断出该句是观点句还是非观点句。

提交格式:

id　　run-tag　　weibo-id　　sentence-id　　opinionated

说明

id：结果序号

run-tag：队伍结果标识

weibo-id：微博 id

sentence-id：句子 id

opinionated：观点句标识，是观点句则为 Y，非观点句则为 N

注：run-tag 的格式为“队伍标识_提交结果组号”，队伍标识可自定，组号用于区分同一队伍的多组提交结果。不同字段之间用\t 隔开。下同。

例如如下两条微博：

weibo1：

<weibo id="1">
<sentence id="1">渭南城管撕春联事件在成都公交车上的分众传媒广泛报道!</sentence>
<sentence id="2">渭南城管真变态啊!</sentence>
</weibo>

weibo2:

<weibo id="2">
<sentence id="1"> #iPad3#这么麻烦的东西怎么还有那么多人在用，又是越狱又是破解。</sentence>
<sentence id="2"> 顺便问一下怎么越狱啊？</sentence>
</weibo>

weibo1 中有两个句子，第一句是非观点句，第二句是观点句。weibo2 中有两个句子，其中第一句是观点句，第二句是非观点句。则正确的输出结果为：

1	xyz	1	1	N
2	xyz	1	2	Y
3	xyz	2	1	Y
4	xyz	2	2	N

注：本评测中观点句的定义不包括表达自我情感、意愿或心情的句子，比如“我感到很高兴”这样的句子是情感句，但不属于本评测定义的观点句。本评测定义的观点句只限定于对特定事物或对象的评价（例如“我真心喜欢 iphone 的屏幕效果。”），不包括内心自我情感、意愿或心情。

评价标准

本任务使用正确率（Precision），召回率（Recall）和 F 值（F-measure）来评价各个参赛队伍对观点句的识别结果。其计算公式如下：

$$\text{Precision} = \frac{\#system\_correct(opinion = Y)}{\#system\_proposed(opinion = Y)}$$

$$\text{Recall} = \frac{\#system\_correct(opinion = Y)}{\#gold(opinion = Y)}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

#gold 是人工标注结果的数目，#system_correct 是提交结果中与人工标注匹配的数目，#system_proposed 是提交结果的数目。

2.2 情感倾向性判断

本任务要求判断微博中每条观点句的情感倾向。评测数据集包含每条微博中的各个句子，参赛队伍需要先进行观点句识别再进行观点句的倾向性分析。观点句的情感倾向可以分为正面（POS），负面（NEG）和其他（OTHER）。

提交格式：

id　　run-tag　　weibo-id　　sentence-id　　polarity

说明

id：结果序号

run-tag：队伍结果标识

weibo-id：微博 id

sentence-id：观点句 id

polarity：情感倾向标识，正面为 POS，负面为 NEG，**中性以及其它无法明确归为正面或者负面的为 OTHER**。

比如：上面 weibo1 和 weibo2 两条微博中，weibo1 的第二句是观点句，情感倾向为负面。weibo2 的第一句是观点句，情感倾向为负面。则其结果应如下：

1　　xyz　　1　　2　　NEG

2　　xyz　　2　　1　　NEG

评价标准

本任务同样使用正确率（Precision），召回率（Recall）和 F 值（F-measure）作为评价标准。

$$\text{Precision} = \frac{\#system\_correct(polarity = POS, NEG, OTHER)}{\#system\_proposed(opinion = Y)}$$

$$\text{Recall} = \frac{\#system\_correct(polarity = POS, NEG, OTHER)}{\#gold(opinion = Y)}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

#gold(opinion = Y)是人工标注结果中观点句的数目，#system_correct(polarity = POS, NEG, OTHER)是提交结果中与人工标注匹配的数目，#system_proposed(opinion = Y)是提交的所有观点句的数目。

3.3 情感要素抽取

本任务要求找出微博中每条观点句作者的评价对象，即情感对象。同时判断针对情感对象的观点极性。评测数据集包含每条微博中的各个句子，参赛队伍需要先进行观点句识别再进行情感要素抽取。

注：

1. 只对微博中的观点句进行情感要素的抽取。
2. **情感对象应首先从当前句子中抽取，如果情感对象不存在于当前句子中，再从整条微博中抽取。对于第二种情况，应优先从当前句子的前一句（包括句子中包含的 hashtag）开始依次向前寻找情感对象，如果没有再从当前句子的后一句（包括句子中包含的 hashtag）开始依次往后寻找。对于那些整条微博中都没有出现的情感对象(有些情况情感对象是隐含的)的观点句，参赛队伍不必进行抽取。**
3. **一个句子中，可以出现多个情感对象，应抽取出每个情感片段所对应的情感对象。**“你根本已经不是个人了，你比蛇还冷血，你比畜生还畜生。”，要求抽取出三个“你”。
4. **抽取情感对象时，要求抽取出尽可能完整和明确的对象，例如“**ipad 的屏幕很棒！**”，要求抽取出情感对象“ipad 的屏幕”，而不仅是“屏幕”。**
5. **对于人称代词（你，我，他，它，你们，我们，他们，它们等）单独作为情感对象出现时，需要在该微博范围内（不包括转发、评论信息）尽量进行指代消解（无法指代消解的情况可以采用这些代词作为对象）**。例如，“小明就读于北京大学，他是名优秀的学生。”情感对象是“小明”而不能是“他”。

提交格式

id run-tag weibo-id sentence-id target begin-offset end-offset polarity 说明

id：结果序号

run-tag：队伍结果标识

weibo-id：微博 id

sentence-id：句子 id

target：情感对象

begin-offset: 情感对象在整条微博中的起始位置

end-offset：情感对象在整条微博中的终止位置

polarity：对情感对象的观点极性，POS 代表正面，NEG 代表负面， OTHER 代表中性或者无法明确归为正面或者负面的其它情形。

注：**对于从当前某个句子中抽取得到的情感对象，其起始位置与终止位置也要基于整条微博来计算。**

比如 weibo1 和 weibo2 的情感要素抽取结果如下：

1	xyz	1	2	渭南城管	26	29	NEG
2	xyz	2	1	iPad3	1	5	NEG

文件采用 unicode（utf-16）编码，每个字符都占两个字节，任意微博中第一个字符的 offset 为 0，第二个字符的 offset 为 1，以此类推。比如：weibo1 第二句开始位置的“渭南城管”这四个字符在整条微博中对应的 offset 分别为 26,27,28,29。评价情感对象时只以 begin-offset 和 end-offset 作为判断依据，target 不参与评价。

评价标准：

本任务同样采用精确（Strict）评价和宽松（Lenient）评价两种方式，均使用准确率（Precision）、召回率（Recall）以及 F 值（F-measure）作为评价标准。

在精确评价中，要求提交的情感对象的 offset 和答案完全相同并且情感对象极性也相同时才算正确。

$$\text{Precision} = \frac{\#system\_correct}{\#system\_proposed}$$

$$\text{Recall} = \frac{\#system\_correct}{\#\text{gold}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

#gold 是人工标注结果中情感对象的数目，#system_correct 是提交结果中与人工标注匹配的数目，#system_proposed 是提交的情感对象的数目。

在宽松评价中，一个结果包含 4 个参与评测的元素：句子微博 id，句子 id，情感对象区间（由起始位置和终止位置构成）和极性，即 *r=(wid, sid, s, p)*。我们首先定义两个结果之间的覆盖率 *c*：

$$c(r,r') = \begin{cases} \dfrac{|s \cap s'|}{|s'|} & if \;\; p = p' \;\&\; wid = wid' \;\&\; sid = sid' \\ 0 & else \end{cases}$$

其中 *s* 和 *s'*为两个结果 *r* 和 *r'*中情感对象的区间，*p* 和 *p'*为对应的极性，*wid* 和 *wid'*为微博 *id*，*sid* 和 *sid'*为句子 *id*。|*| 表示计算区间的长度。

两个结果集合 *R* 和 *R'*之间的覆盖率 *C* 定义为：

$$C(R,R') = \sum_{r_i \in R} \sum_{r_j' \in R'} c(r_i, r_j')$$

假设提交的结果集合为 *R'*，标注结果集合为 *R*，则精度、召回率和 F 值为：

$$\text{Precision} = \frac{C(R,R')}{|R'|}$$

$$\text{Recall} = \frac{C(R',R)}{|R|}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中|*|表示计算集合中元素的个数。

4 数据集

本次评测数据来自腾讯微博 [1]。评测数据全集包括 20 个话题，每个话题采集大约 1000 条微博，共约 20000 条微博。数据采用xml格式，已经预先切分好句子。

数据样例：

```
<weibo id="1">
```

[1] http://t.qq.com/

<sentence id="1">#广外神仙姐姐# 近广东外语外贸大学南国商学院的一位名叫林雪薇的女老师火了，她的火是因为她的美丽。</sentence>
<sentence id="2">天生丽质的林雪薇被网友称为“神仙姐姐”。</sentence>
<sentence id="3">据说，她的课几乎没人逃课。</sentence>
<sentence id="4">“何止逃课人数减少，当年许多不是她的课的学生都说要去上她的英语课。</sentence>
<sentence id="5">想要火不一定要穿的少，漂亮就好，http://url.cn/0g1mXC</sentence>
<hashtag id="1">广外神仙姐姐</hashtag>
<forward id="1">/玫瑰 || @s1144191251: 亲爱的，要记得：你不坏，你不赖，你不差，你很棒。#广外神仙姐姐# /玫瑰</forward>
<forward id="2">亲爱的，要记得：你不坏，你不赖，你不差，你很棒。#广外神仙姐姐# /玫瑰</forward>
< comment id="1">确实不错啊 </ comment >
</weibo>

其中，每条微博对应一个<*weibo*>元素，每个句子对应一个<*sentence*>元素，每条微博的hashtag、转发和评论分别对应标签<*hashtag*>，<*forward*>和<*comment*>。注：**只有部分微博具有转发和评论信息。由于 hashtag 跟微博文本之间难以明确区分，本次评测在微博文本中保留 hashtag 信息，对微博分句之后 hashtag 信息会附属在最邻近的句子上。评论和转发信息仅作为辅助信息出现，不参与评测，参赛队伍可以自由选择使用或者不使用这些信息。**

所有数据文件均采用 unicode（utf-16）编码。**读取数据集时，需要进行 XML 实体和字符间的转换，主要包含以下 5 组：(<，<)，(>，>)，(&，&)，('，')，("，")。如果不做转换，会影响情感要素抽取任务中情感对象的位置区间。**

5 评测方法

本次评测为离线评测。参评单位自行处理数据，按照规定格式生成相应结果后提交。答

案采用人工标注的方法确定。参赛单位需要处理全部评测数据，但用于实际评测的人工标注数据仅为评测数据全集的 10%左右。

具体评测步骤为：

1) 评测单位预先提供测试样例（包括答案）；
2) 评测单位给出测试数据；
3) 参评单位运行被测系统，得出测试结果；
4) 参评单位提交测试结果；
5) 评测单位标注答案，运行自动评测程序，统计评测结果；

6 评测要求

参评单位应当采用自动的方法，针对微博进行情感分析。参评系统应当预先训练模型、调整好所有参数，运行过程中不得有人工干预。本次评测不限制使用各种语义资源。对于每个子任务，参评单位至多提交 2 组结果。

7 评测日程

2012/1/1-2/29： 起草评测大纲，征求各方意见；

2012/3/1-3/31： 修订完善评测大纲，确定评测数据；

2012/4/1： 发布评测任务，接受评测报名；

2012/5/4： 发布评测样例数据集；

2012/5/31： 评测报名截止；

2012/5/1-6/30： 构建评测数据集，制定标准答案；

2012/7/1： 发布评测数据集；

2012/7/31： 参评单位提交运行结果；

2012/8/1-8/31： 组织专家评测小组进行结果评判，发布评测结果；

2012/9/1-9/24： 征集评测论文；

2012/9/25-9/30： 确定受邀报告；

2012/10/29-11/6： 宣读报告，交流经验和技术；

8 如何注册

参加评测的单位需要在接受报名的时间内到如下评测主页下载报名表，并在截止日期前通过电子邮件或传真方式发送给评测组织者。报名应以研究机构或公司为单位,暂不接收个人报名。

http://tcci.ccf.org.cn/conference/2012/

如果你有任何关于本次评测的问题请发邮件至：huangxiaojiang@pku.edu.cn

9 本次评测的组织

- **主办单位**

  中国计算机学会(CCF)

- **承办单位**

  北京大学

  MSRA

- **协办单位**

  数字出版国家重点实验室

- **评测委员会（按照姓氏拼音排序）**

  李寿山、刘群、万小军、韦福如、吴云芳、徐睿峰