Filtered Local Linear Embedding Classifier for Stance Detection

# 1.Introduction

在社交网络中，如Twitter、Facebook，伴随着一个新的话题热潮，针对该话题的讨论会在短时间内大面积的爆发。而这些评论性的文本往往带有用户个人的情感（positive or negative）与立场（favor or against）。sentiment analysis 的研究已经有大量的研究结果，而近期，stance detection成为Semantic Analysis领域的新的研究课题。在语义评测会议SemEval 2016中，发布了针对五个热门话题（"Atheism", "Climate Change is a Real Concern", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion"

）的Stance Detection 任务。对于每一个Tweet，对应“是否与话题相关”与“评论的立场”，将每个Tweet分为None，Favor与Against。

stance detection不同于sentiment analysis。首先，需要判断文本是否与话题有关，话题的相关性与文本附带的情感并没有直接关系。其次，情感与立场并不完全一致，negative不代表against，favor也不代表positive。比如对于无神论的话题，对God的positive与对无神论的positive分别代表着Against与Favor，这就意味着许多基于sentiment word的方法不能直接有效作用于stance detection，包括大量以情感词作为特征的机器学习方法。

对于判断话题相关度，我们设计Topic-keyword的方法用于检测该Tweet是否与话题相关；对于立场分析，我们参照传统LLE算法，并且在其基础上，使用机器学习的方法，设计了Filtered LLE（FLLE）方法，并对参数的选择进行自动化。我们将用FLLE方法与SLLE、深度学习方法进行对比。最后，我们将在SemEval Task6的数据集上讨论我们的结果与其他方法的对比。

## 2.Related Work

Sentient analysis

## 3.Topic Keyword Detection

初步定为对每一个话题，通过Wiki检索，将Topk的文档作为语料集，然后利用数据集合中的正例（Favor&Against）与负例（None）来筛选Topic Keyword List。然后判定None基于一个假设：针对该话题的文本包含且至少包含一个Topic Keyword。

## 4.Filtered LLE

4.1 LLE

4.2 SLLE

4.3 FLLE

LLE的方法无法运用训练集的标签，而SLLE仅仅在训练集数据的降纬中使用了标签，强制将不同种类的数据分开来。

一个Tweet的立场判断仅能从最相似的几个Tweet得到较为准确的信息，这正好与LLE的局部性吻合，因此我们有理由相信LLE有适用于Stance Detection任务的特性。

然而在Stance Detection的任务中，因为文本相似度并不能反映立场相似度。两个文本相似的Tweet，极有可能恰好是相反立场（**例子与数据统计，有50%左右的Tweet与其相似度Top5的Tweet立场不同**），这是由于Tweet的140词限制，属于短文本，所以少量词的相似就可能造成两个不同立场Tweet相关度远高于同一立场的两个Tweet。因此两种LLE的方法又都是不适用的。

因此，我们设计了FLLE的方法，通过机器学习的方法，为每一个数据的重构集提供一个Filter，从而保证尽可能一个数据被其同类数据所重构。方法描述如下：

1. 所有Tweet分为训练集A与测试集B。
2. 对于A中的每一个Tweet a，选择A中与其最相似的Top 5 Tweet {a1,…,a5}，将a分别与这5个Tweet的0-1词向量的异或向量作为训练数据，同类标为1，异类标为0
3. 使用机器学习方法进行Predict模型训练。
4. 对于A+B中的所有Tweet分别与A中的所有Tweet构成0-1异或向量，利用3的模型进行预测，预测结果记为P，P为(|A|+|B|)\*|B|的矩阵。
5. 使用LLE方法对A中所有Tweet进行映射，区别在于，对于每一个Tweet i，仅在P的第i行中前k大的Tweet中选择。其他的同LLE方法一致。
6. 对于B中所有Tweet，同样，对于每一个Tweet i，仅在P的第i行中前k大的Tweet中选择。然后使用步骤5中的映射结果进行重构。
7. 对最后得到的降维结果使用分类方法进行训练与预测。

在步骤2中的测试集的构成上，限制了两个有一定内容相关度的Tweet才会被机器学习方法所学习，使得毫无关系的两个Tweet不会被考虑。同时LLE的重构集选择方法也保证了最终重构集会与数据有足够的内容相似度，达成了优美的一致。最后使用KNN分类器进行分类。

相较于LLE与SLLE方法，FLLE使用Filter的方法，保障的训练集与测试集在重构集的选择上的一致性，同时充分运用到Tweet与Tweet之间的信息。

## 5.Improvements to Filtered LLE

5.1 Automatic Filter parameter selection（Filter的k参数）

5.2 KNN method for classification

|  |  |  |  |
| --- | --- | --- | --- |
| Topic | Train数 | 正确数 | K |
| Feminist Movement | 538 | 522 | 143 |
| Atheism | 396 | 376 | 19 |
| Hillary Clinton | 473 | 445 | 17 |
| Legalization of Abortion | 439 | 429 | 29 |
| Climate Change is a Real Concern | 227 | 215 | 7 |

5.3 Automatic KNN parameter selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Train | BadSelect | Test | BadSelect |
| Feminist Movement | | 538 | 13 | 285 | 91 |
| Atheism | | 396 | 41 | 220 | 61 |
| Hillary Clinton | | 473 | 47 | 295 | 119 |
| Legalization of Abortion | | 439 | 20 | 280 | 84 |
| Climate Change is a Real Concern | | 227 | 15 | 169 | 46 |

## 6.Experiment

6.1 Dataset

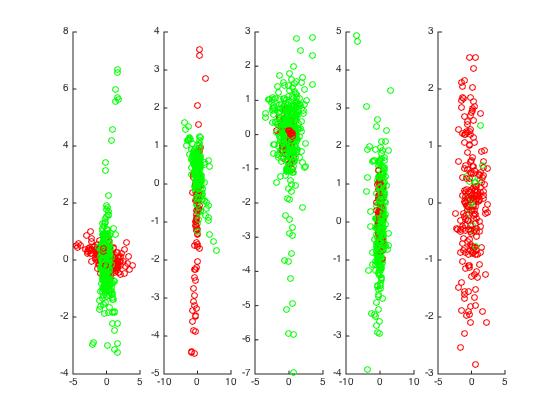
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Train | | Test | |
|  |  | FAVOR | AGAINST | FAVOR | AGAINST |
| Feminist Movement | | 210 | 328 | 58 | 227 |
| Atheism | | 92 | 304 | 32 | 188 |
| Hillary Clinton | | 112 | 361 | 45 | 250 |
| Legalization of Abortion | | 105 | 334 | 46 | 234 |
| Climate Change is a Real Concern | | 212 | 15 | 123 | 46 |

数据分布不均匀

6.2 Result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| topic | stance | precision | recall | f-score | **(A\_F+F\_F)/2** |
| Atheism | FAVOR | 0.3333 | 0.2188 | 0.2642 | 0.5443 |
| AGAINST | 0.7437 | 0.925 | 0.8245 |
| Climate | FAVOR | 0.7246 | 0.9837 | 0.8345 | 0.4172 |
| AGAINST | 0 | 0 | 0 |
| Feminist | FAVOR | 0.375 | 0.5172 | 0.4348 | 0.5937 |
| AGAINST | 0.7122 | 0.7978 | 0.7526 |
| Hillary | FAVOR | 0.5357 | 0.3333 | 0.411 | 0.5722 |
| AGAINST | 0.603 | 0.936 | 0.7335 |
| **Abortion** | FAVOR | 0.4872 | 0.413 | 0.4471 | 0.6259 |
| AGAINST | 0.7178 | 0.9153 | 0.8047 |
| All | FAVOR | 0.5731 | 0.6316 | 0.6009 | 0.686 |
| AGAINST | 0.6871 | 0.8783 | 0.771 |

6.2.1 LLE，SLLE，FLLE的比较与描图



6.2.2 SVM、KNN与KNN auto-select比较

6.2.3 FLLE与Deep Learning方法（Word2Vector、Para2Vector的比较）的精确度与效率比较

6.2.4 FLLE与SemEval其他组的比较

|  |  |
| --- | --- |
| Team Name | Official Metric |
| FLLE | 0.686 |
| MITRE | 0.6782 |
| pkudblab | 0.6733 |
| TakeLab | 0.6683 |
| PKULCWM | 0.6576 |
| ECNU | 0.6555 |

FLLE是简洁，高效并且适用于短文本分类与立场分析的分类方法，并且在数据分布相当不均匀的情况下依然取得良好的效果。

## 7.Future Work

FLLE是通用的分类方法，其相较于深度学习的高效与良好的准确性，值得在大量文本分类的问题中进行研究。