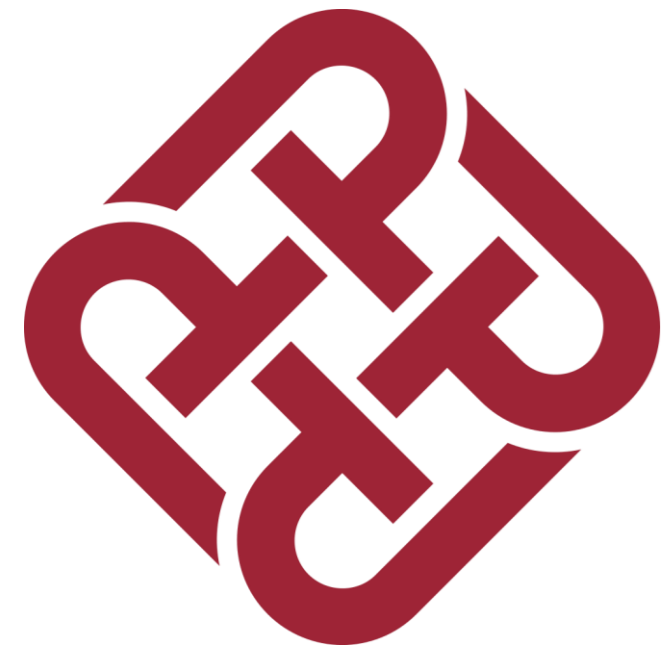


Emotion in Code-switching Texts: Corpus Construction and Analysis



Sophia Yat Mei Lee¹ and Zhongqing Wang^{1,2}

¹Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

²Natural Language Processing Lab, Soochow University, Suzhou, China

{sophiaym, wangzq.antony}@gmail.com



Objectives

- To provide a well-defined and efficient method for **constructing** and **analyzing** a **Chinese-English code-switching corpus** from social media
- To propose an annotation scheme to identify the emotions associated with the languages expressing them
- To propose a **multiple-classifier-based automatic detection approach** to detect emotions in code-switching text

Data Collection and Annotation

- 4,195 code-switching posts are extracted from **Weibo.com**
- Five basic emotions are annotated, namely **happiness, sadness, fear, anger and surprise**
- 4 kinds of caused situations are identified, i.e. **English, Chinese, Both, Mixed**

Examples of Code-switching Emotional Posts

[E1] 静静坐下来看别人**show**啦。刚刚在节目里看到妈咪和弟的视频真的很意外！ (**Surprise, CN**)

(*I set down quietly to watch someone else's **show**. To my surprise, both my mother and brother appeared on the programme!*)

[E2] 玩了一下午轮滑 **so happy** ! (**Happiness, EN**)

(*I went rollerblading the whole afternoon, **so happy**!*)

[E3] 开学以来，浮躁的情绪。不安稳的心态。确实该自己检讨一下了。。。 **sigh**~~~ (**Sadness, BOTH**)

(*I have been grumpy and emotional since the first day of school, unstable mindset too. It's really time to self-evaluate...**sigh**~~~)*)

[E4] 上了一天的课，嗓子**hold**不住了啊！ (**Sadness, MIXED**)

(*I have been teaching the whole day, my throat **can't take it anymore**!*)

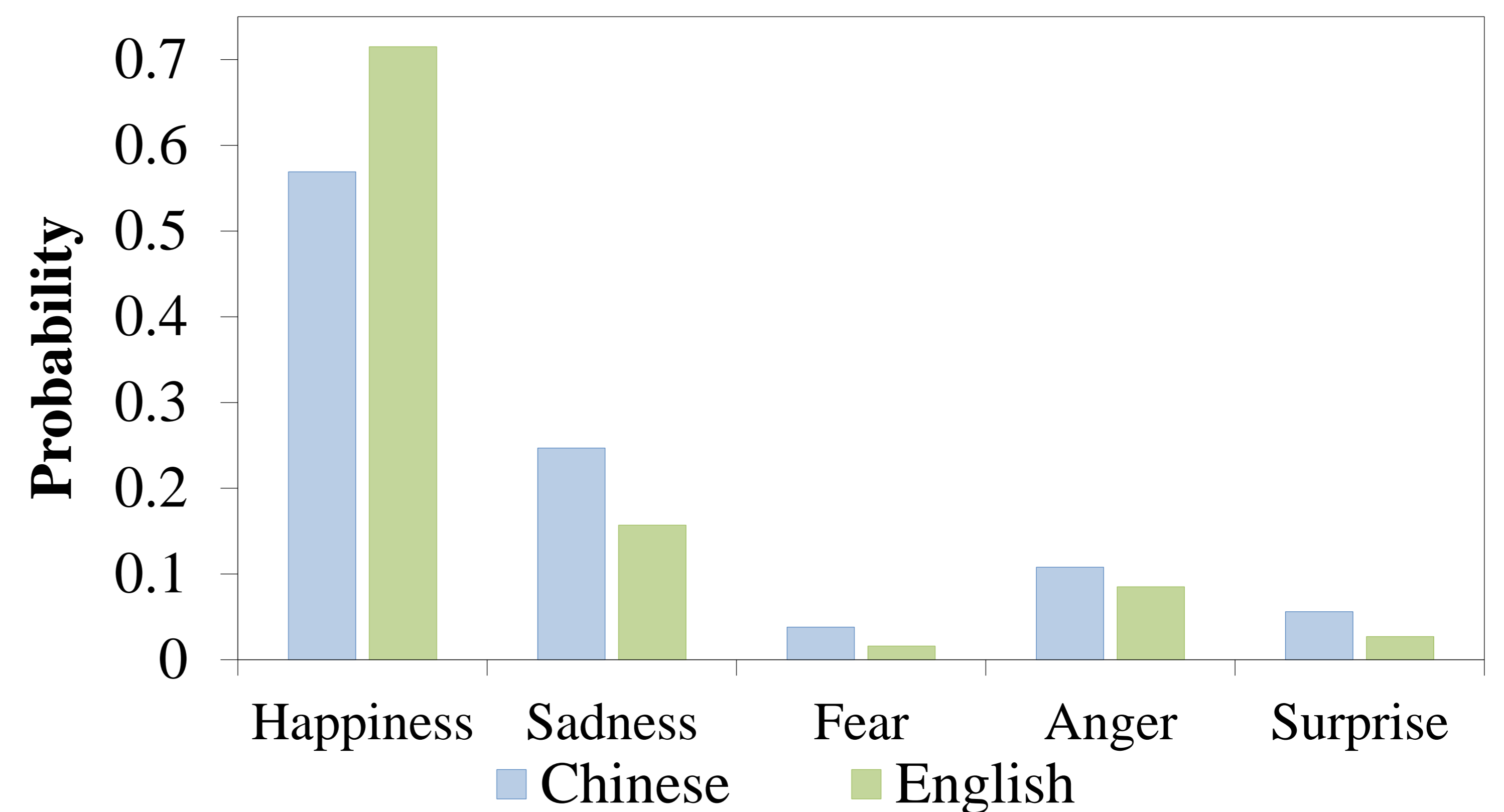
Statistics and Analysis

Agreement Analysis

	Kappa Score
Emotion	0.692
Caused Language	0.767

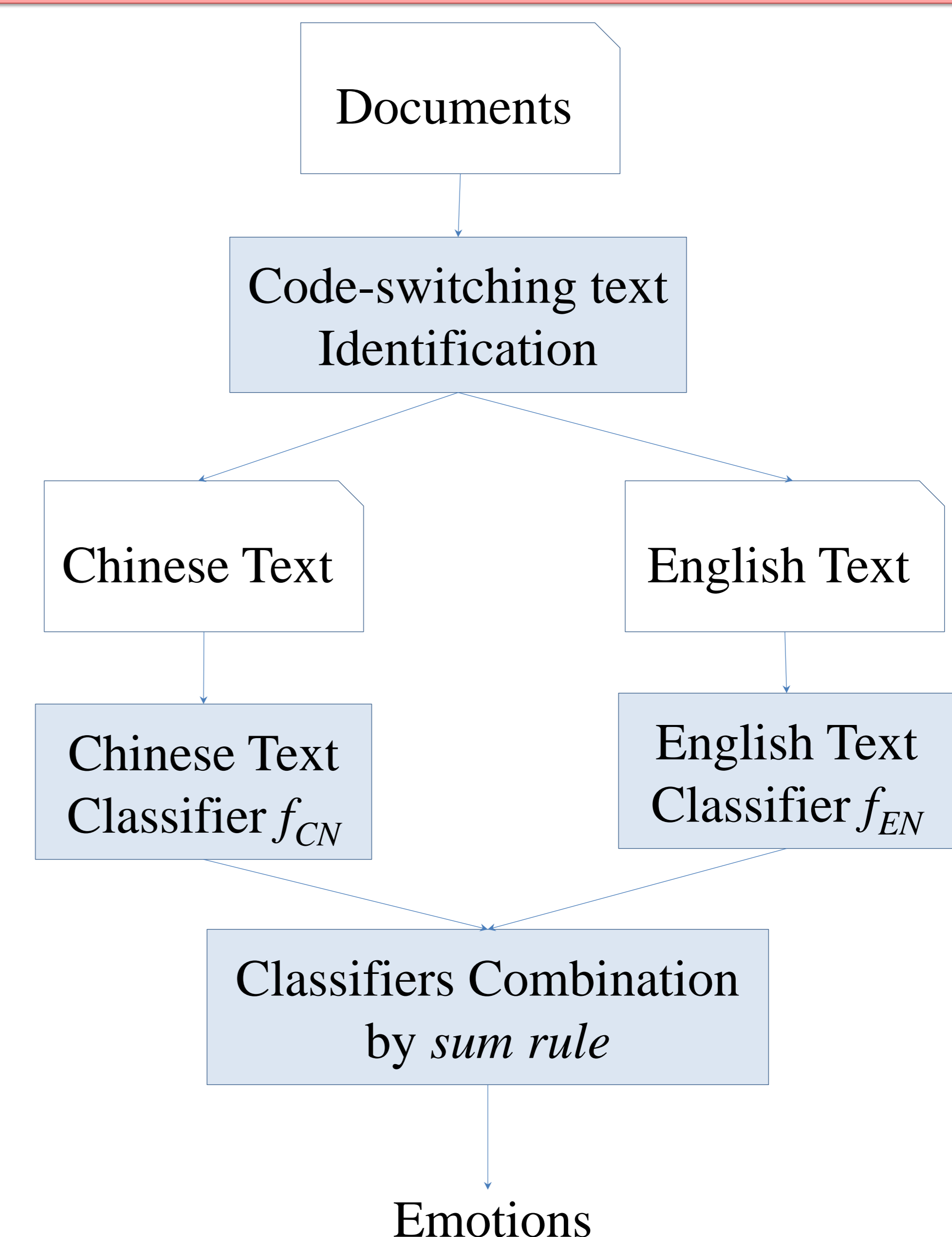
General Distribution of Data

- Out of 4,195 posts, 2,312 are of emotions
- 81.4% of emotions are expressed in **Chinese**, 43.5% are in **English**
- 33.1% of the posts are concerned with **Life**, followed by **Service** (26.3%), **Products** (20.4%) and **Celebrities** (11.4%)



Joint Distribution of Emotions and Caused Languages

Automatic Emotion Detection in Code-switching Texts



Overview of the Multiple-classifiers-based Detection Framework

Experiment Results

- Half of the annotated posts are selected as the training data and another *half* as test data
- FudanNLP** is adopted for Chinese word segmentation and Maximum Entropy (ME) as the basic supervised classification model, while the ME algorithm is implemented with the **MALLET Toolkit**

	Acc.
f_{ALL}	0.509
f_{CN}	0.521
f_{EN}	0.409
f_{comb}	0.539

Results of Emotion Detection in Code-switching Text