sampling where the active learner selects those unlabeled samples which have the largest disagreement among several committee classifiers. Besides query by committee (QBC) as the first of such type (Freund et al., 1997), co-testing learns a committee of member classifiers from different views and selects those contention points (i.e., unlabeled examples on which the views predict different labels) for manual annotation (Muslea et al., 2006).

However, most previous studies focus on the scenario of balanced class distribution and only a few recent studies address the active learning issue on imbalanced classification problems including Yang and Ma (2010), Zhu and Hovy (2007), Ertekin et al. (2007a) and Ertekin et al. (2007b)[2]. Unfortunately, they straightly adopt the uncertainty sampling as the active selection strategy to address active learning in imbalanced classification, which completely ignores the class imbalance problem in the selected samples.

Attenberg and Provost (2010) highlights the importance of selecting samples by considering the proportion of the classes. Their simulation experiment on text categorization confirms that selecting class-balanced samples is more important than traditional active selection strategies like uncertainty. However, the proposed experiment is simulated and non real strategy is proposed to balance the class distribution of the selected samples.

Doyle et al. (2011) propose a real strategy to select balanced samples. They first select a set of uncertainty samples and then randomly select balanced samples from the uncertainty-sample set. However, the classifier used for selecting balanced samples is the same as the one for supervising uncertainty, which makes the balance control unreliable (the selected uncertainty samples take very low confidences which are unreliable to correctly predict the class label for controlling the balance). Different from their study, our approach possesses two merits: First, two feature subspace classifiers are trained to finely integrate the certainty and uncertainty measurements. Second, the *MA* samples are automatically annotated, which reduces the annotation cost in a further effort.

3 Active Learning for Imbalanced Sentiment Classification

Generally, active learning can be either stream-based or pool-based (Sassano, 2002). The main difference between the two is that the former scans through the data sequentially and selects informative samples individually, whereas the latter evaluates and ranks the entire collection before selecting most informative samples at batch. As a large collection of samples can easily gathered once in sentiment classification, pool-based active learning is adopted in this study.

Figure 1 illustrates a standard pool-based active learning approach, where the most important issue is the sampling strategy, which evaluates the informativeness of one sample.

```
Input:
    Labeled data $L$;
    Unlabeled pool $U$;
Output:
  New Labeled data $L$
Procedure:
Loop for $N$ iterations:
(1). Learn a classifier using current $L$
(2). Use current classifier to label all unlabeled
     samples
(3). Use the sampling strategy to select $n$ most
     informative samples for manual annotation
(4). Move newly-labeled samples from $U$ to $L$
```

Figure 1: Pool-based active learning

3.1 Sampling Strategy: Uncertainty vs. Certainty

As one of the most popular selection strategies in active learning, uncertainty sampling depends on an uncertainty measurement to select informative samples. Since sentiment classification is a binary classification problem, the uncertainty measurement of a document d can be simply defined as follows:

$$Uncer(d) = \min_{y \in \{pos,neg\}} P(y \mid d)$$

Where $P(y \mid d)$ denotes the posterior probability of the document d belonging to the class y and {pos,

[2] Ertekin et al. (2007a) and Ertekin et al. (2007b) select samples closest to the hyperplane provided by the SVM classifier (within the margin). Their strategy can be seen as a special case of uncertainty sampling.