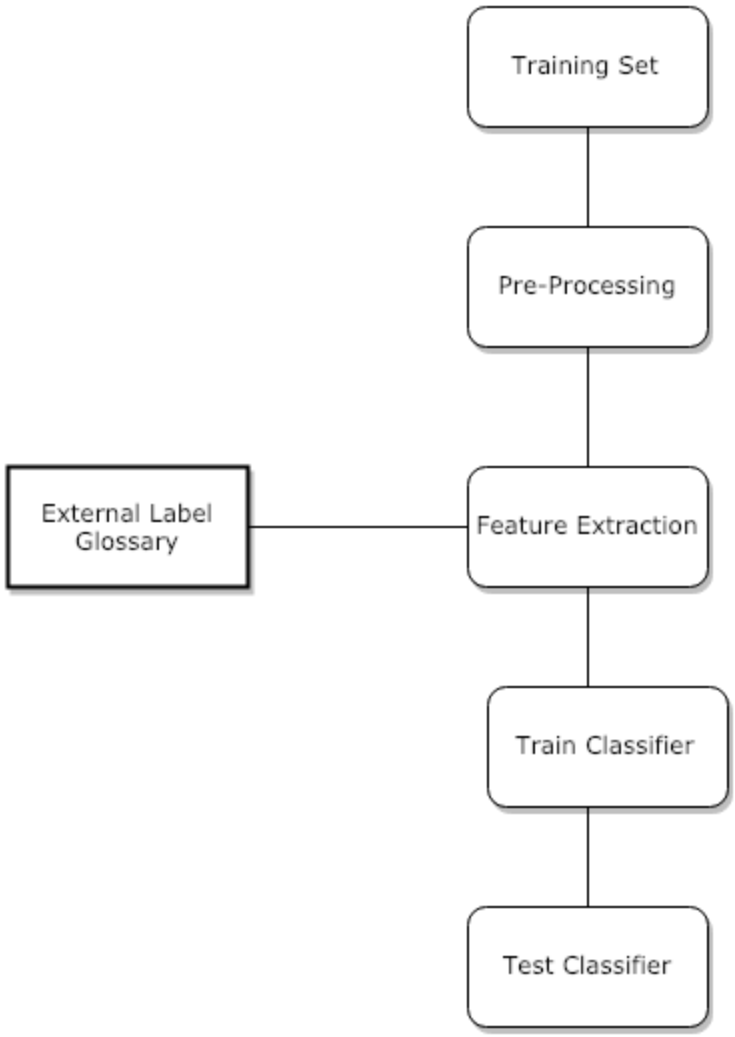
Automatic text classification of sports blog data

Ben Stewart S 2016103513
Siva Kailash S 2016103593

Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given text document to a set of predefined categories based on the features extracted from its textual content. We will be attempting to automatically classify the textual entries made by bloggers on various topics, to the appropriate category by following steps like preprocessing, feature extraction and naïve **Bayesian classification**. **Empirical evaluation** must result in a very high accuracy. In addition to classifying the textual entries of blogs, it is proposed that the extracted features themselves be further classified under more meaningful heads which results in generation of a semantic resource that lends greater understanding to the classification task. This semantic resource can be used for data mining requirements that arise in the future.

We taken the paper that classifies text of sports blog data that works only with sports data. It is limited to very small tagset. Here our aim is to study general blog data and classify them. The process we are going to do is given in the following Block diagram.

BLOCK DIAGRAM :



DATASETS :

The training datasets are taken from https://www.kaggle.com/rtatman/blog-authorship-corpus.

The Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words - or approximately 35 posts and 7250 words per person.

All bloggers included in the corpus fall into one of three age groups:

- 8240 "10s" blogs (ages 13-17),
- 8086 "20s" blogs(ages 23-27)
- 2994 "30s" blogs (ages 33-47).

For each age group there are an equal number of male and female bloggers.