PAN @ CLEF 2013

Automatic Author Profiling Based on Linguistic and Stylistic Features

Braja Gopal Patra[1], Somnath Banerjee[1], Dipankar Das[2], Tanik Saikh[1], Sivaji Bandyopadhyay[1]
Department of Computer Science and Engineering, Jadavpur University, Kolkata, India[1]
Department of Computer Science & Engineering, NIT Meghalaya, India[2]

Overview in Details

Smiley words list: Frequency of Smileys in each file has been calculated using the handcrafted Smiley List consisting of 55 smileys. After calculating the frequency of smileys in each of the files, smileys are replaced by full stop words.

Word class Frequency: Each word class contains a set of stemmed words related to synonyms and hypernyms. We have used the hypernym and synonym relations of RiTaWordNet to increase the seed list. There are 9 classes, namely money, job, sports, television, sleep, eat, sex, family and friend.

Positive and Negative word class : These two classes contain the words which do not appear in our existing 9 word classes. After getting all possible POS from RiTaWordNet, the sentiment scores of the words have been calculated using the SentiWordNet 3:0 lexicon. Threshold value: |0.1|.

Stop words frequency: We have observed that the age group 20 has used more number of stop words in their text. A total of 329 stop words have been prepared manually.

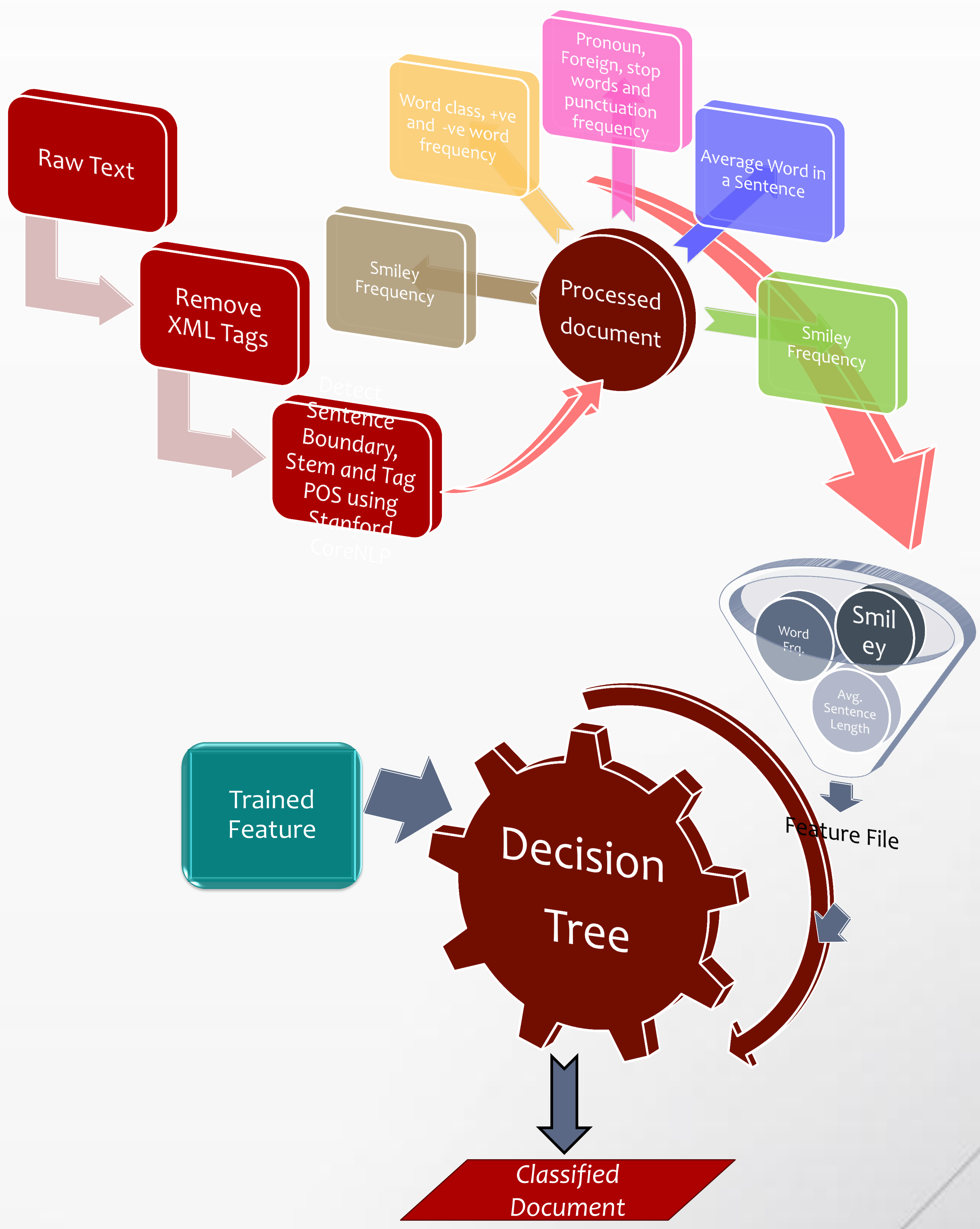
List of Foreign Words (FW): These are the words, which are tagged as FW by the StanfordCoreNLP POS tagger. These are basically *meee*, *yesss*, *thy*, *u* and *urs* etc.

List of Punctuations: 10 types of punctuations are prepared manually.
List of Pronouns: The frequencies of the pronouns are also computed. Pronouns are tagged as PRP by StafordCoreNLP POS tagger.

Average Length of Sentence: We have considered the average sentence length in documents. The sentence boundary is detected by the StanfordCoreNLP tool.

It has been found that the size of each document varies, i.e., some documents contain more number of words and some documents contain less words. So, we have normalized each bag of word feature by dividing the total number of words in a document.

System Architecture



Statistics of Word class

Class	Number of Words
Money	881
Job	1145
Friends	508
Family	302
Eating	3120
TV	261
Sports	591
Sleep	19
Sex	1008
Positive	9627
Negative	10383

Results

Gender	Age	Overall
56.83	28.95	15.74

Conclusion

- This work is of interest for a number of potential applications like forensics, security and marketing etc.
- Same template used for both gender and age classification and this may be one of the reason for degradation in age classification.
- In our future work, the accuracy of the classification can be improved by finding and incorporating more suitable features like POS, number of Ellipsis, average word length and number of paragraphs etc.
- It would also be interesting to perform deeper features engineering for finding demographic and psychometric author traits more correctly.

Reference

- Houvardas, J., and Stamatatos, E.: N-gram feature selection for authorship identification. Artificial Intelligence: Methodology, Systems, and Applications. Springer Berlin Heidelberg, 77-86 (2006).
- Schler, J., Koppel, M., Argamon, S., and Pennebaker, J.: Effects of age and gender on blogging. In Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 199-205 (2006).
- Koppel, M., Argamon, S., and Shimoni, A. R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412 (2002).

The work has been carried out with support from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.