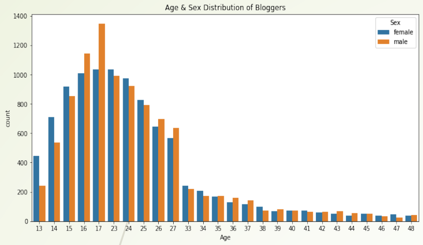
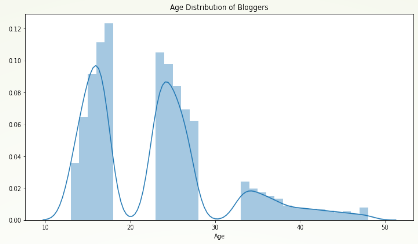
Blog Posts Analysis

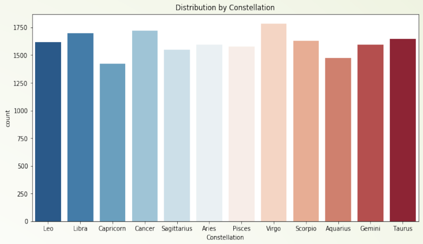
Guoqiang Liang

Introduction

The Blog Authorship 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. Each blog is associated with blogger's self-provided gender, age, industry and astrological sign. All are labeled for gender and age but for many, industry and/or sign is marked as unknown.

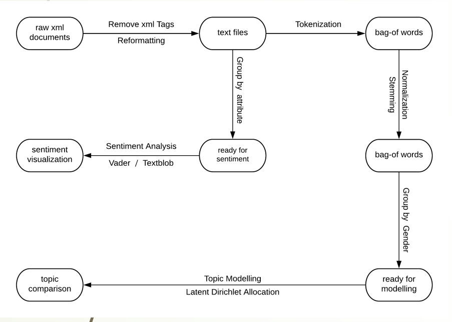




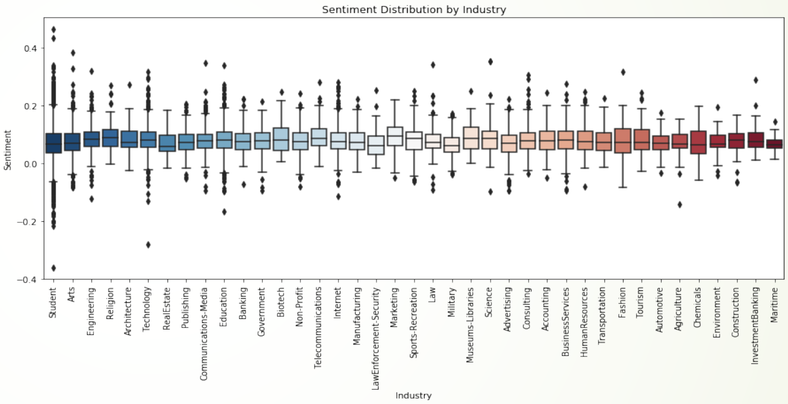


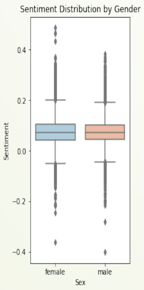
Given the dataset, we can conduct sentiment analysis on the blogs with respect to different gender, age group, constellation and industry. Also, Latent Dirichlet Allocation will be used to model the topics that male and female care about the most respectively.

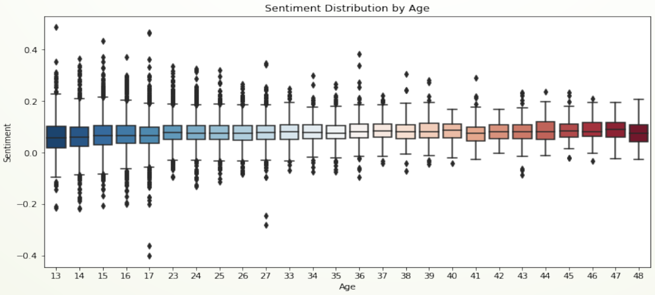
Pipeline

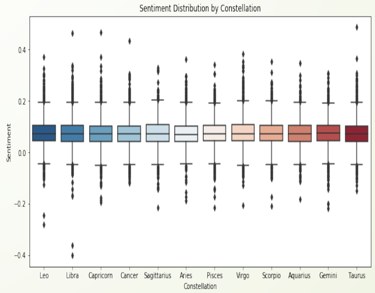


Sentiment Analysis









Topic Modeling

	Topics – Male Bloggers					Topics – Female Bloggers				
1	site	information	software	http	computer	haha	home	dun	den	coz
2	time	day	night	work	life	night	home	week	morning	work
3	way	school	today	love	home	today	night	fun	school	home
4	war	government	world	life	president	world	book	life	church	story
5	que	mobynath	pero	ako	lang	kung	lang	pero	hindi	kasi