در این پروژه ابتدا دو فایل ورودی خوانده شده و سپس بعد از نرمال سازی و sentence tokenizing داده ها را به فرمتی که کاراکتر اول در هر خط شماره خط و کاراکتر دوم شماره کلاس که از دو عدد ۱ و -۱ استفاده شده و بقیه کاراکتر ها به عنوان feature در نظر گرفته می شوند که در اینجا کل جمله به عنوان feature در نظر گرفته شده و حالت feature های نامحدود پیاده سازی شده است.

برای قسمت feature های محدود feature خاصی مد نظر نبود و قسمت POS tagging در هضم error داشت و قابل استفاده نبود.

برای استفاده از mallet از دو دستور زیر استفاده می کنیم:

Bin/mallet import-file –input result.txt –output data.mallet

در اینجا ورودی با فرمت آماده شده result.txt و خروجی به صورت فایل mallet با نام data.mallet‌ است.

Bin/mallet train-classifier –input data.mallet –cross-validation 10 –trainer MaxEnt –trainer NaiveBayes

