گزارش مربوط به کلاسه‌بندی پروژه:

مقایسه‌ی فالوئرهای دو فرد شاخص در توییتر

قسمت اول،

برای تمیز کردن ابتدا تمام لینک‌های توئیت را حذف می‌کنیم سپس تمامی کارکترهای انگلیسی را نیز پاک می‌کنیم. بعد از آن منشن‌ها را نیز از بین می‌بریم و سپس تمام کارکترهای عربی را جایگزین می‌کنیم. برای اعداد هم تمامی ارقام (از جمله فارسی و عربی) را با معادل انگلیسی آن جایگزین می‌کنیم.

در مرحله‌ی بعد تمامی علامت‌های جمع «ها» را با کلمه‌ی‌قبل + نیم‌فاصله + ها جایگرین می‌کنیم. این کار را برای تمامی «می» استمراری هم انجام می‌دهیم.

برای سایر نیم‌فاصله‌ها هم آن‌ها را حذف می‌کنیم تا تمام کلمات ما یک‌دست شود.

سپس برای هر کدام لیستی از کلمات اضافه و فعل‌های بی‌معنی تشکیل داده و این کلمات را از دیتاست حذف می‌کنیم.

برای هر توئیت هم دقت می‌کنیم که حتما زبان آن فارسی باشد. (این فیلد توسط خود توئیتر فراهم شده‌است.)

ضمنا برای هشتگ‌ها ابتدا علامت # را از بین می‌بریم و سپس تمامی _ ها را با نیم‌فاصله جایگزین می‌کنیم.

جدا سازی جملات و توکن‌ها: در این قسمت از ابزار hazm استفاده می‌کنیم که دقت مناسبی در جداسازی برای ما به ارمغان می‌آورد. سپس علامت‌هایی مثل نقطه، ویرگول، پرانتز باز/بسته، علامت سوال و ... را از بین توکن‌ها حذف می‌کنیم

نمونه‌ی یک توییت بعد از انجام همه‌ی مراحل:

```
اسلام قرآن قبول حجاب قبول نمیزاره اسلام
```

جداسازی دیتای آموزش و تست: کل دیتاست شامل 2133129 توییت را به دو قسمت تقسیم می‌کنیم، قسمت اول شامل ۸۰ درصد دیتا برای آموزش و ۲۰ درصد باقی برای تست.

نکته‌ی قابل توجه این است که توزیع کلاس‌ها در هر قسمت با هم برابر است یعنی اگر نسبت کلاس اول به کلاس دوم ۴ به ۱ باشد، این نسبت در هر بخش از دیتا رعایت شده است.

گزارش نتایج:

از آن‌جایی که در این مسئله برچسب «درست» یا «غلط» وجود ندارد و بین کلاس‌ها هیچ برتری وجود ندارد ما نتایج را برای هر کلاس به صورت جداگانه محاسبه می‌کنیم.

معیار / برچسب	برچسب «اکبرین»	برچسب «رسایی»
accuracy	88.67	88.67
precision	88.64	78.50
recall	94.45	62.68
f1	91.46	69.65

موثر ترین کلمات:

برای به دست آوردن این کلمات به ازاری هر کلمه احتمالش را در هر دو کلاس حساب می‌کنیم و سپس آن‌ها را از هم کم می‌کنیم و سپس اختلاف‌ها را نزولی مرتب می‌کنیم:

برچسب «اکبرین»	برچسب «رسایی»
حال	حرف
کاش	بنده
باز	جناب
بدن	علیه
رفتن	قرار
کردید	وجود

برچسب «اکبرین»	برچسب «رسایی»
ببینید	زدن
بدید	قبول
فعلا	دنبال
میشود	عده

قسمت دوم،

آماده سازی دیتا برای وپال‌وبیت همانند قبل است و فرق چندانی ندارد.

دستور مورد استفاده برای آموزش:

```
$ vw -d data_set_train.vw -c --passes 10 -f predictor.vw –ngram <n>
––loss <loss>
```

دستور مورد استفاده برای تست:

```
$ vw -d data_set_test.vw -t -i predictor.vw -p predictions.txt
```

hinge loss:

معیار / برچسب	برچسب «اکبرین»	برچسب «رسایی»
accuracy	74.97	74.97
precision	95.34	85.98
recall	49.98	99.20

logistic loss 1-gram:

معیار / برچسب	برچسب «اکبرین»	برچسب «رسایی»
accuracy	87.49	87.49
precision	88.96	87.23
recall	55.75	97.76

logistic loss 2-gram:

معیار / برچسب	برچسب «اکبرین»	برچسب «رسایی»
accuracy	86.81	86.81
precision	86.14	86.93
recall	54.87	97.14

logistic loss 3-gram:

معیار / برچسب	برچسب «اکبرین»	برچسب «رسایی»
accuracy	86.16	86.16
precision	83.88	86.59
recall	53.72	96.66

با توجه به نتایج استفاده از n-gramها اعداد بدتری به ما می‌دهند و لاجیستیک جواب بهتری برای این تسک است.