گزارش Navie Bayes Classifier

محمد فرهی شماره دانشجویی: ۸۱۰۱۹۸۴۵۱

1- هر دو روش stemming و lemmatization، یک روش normalization به حساب می آیند و برای از بین بردن فرم اشتقاقی یا فرم مخصوصی از یک کلمه در یک متن به کار می رود. در stemming به صورت نابخردانه، معمولا چند حرف آخر کلمه ها را حذف می کند که معمولا باعث از بین رفتن پسوند ها می شود (کتاب ها -> کتاب) و بعضی اوقات هم باعث بی معنی شدن کلمه می شود(برای ->برا). در روش lemmatization، سعی می شود ریشه کلمات به جای آن فرم خاصی از کلمه که مخصوص آن متن است جایگزین شود. (می روم ->رو#رفت) در این بخش از هر دو استفاده شده است. ابتدا به متن به کلمات تشکیل دهنده شده است و سپس stemming و lemmatization بر رو آن ها انجام می شود. در آخر سعی می شود stop word هایی مثل حروف اضافه و کلمات تک کاراکتری (که معمولا علائم نگرشی هستند) که تاثیری روی پیشبینی ندارند حذف شوند.

2- evidence یا predictor prior : همان شواهدی است که بر اساس آن احتمال شرطی postier را برای متغیر تصادفی هدف را حساب می کنیم(در این جا جمله یا لیست کلمات). این احتمال(P(x را با استفاده از قانون احتمال کل به دست می آوریم.
likelihood : احتمال رخداد جمله یا متن x به شرط این که مربوط به کلاس c باشد. برای محاسبه این احتمال، به صورت bag of

words عمل می کنیم. به این صورت که ابتدا احتمال رخداد یک کلمه را به شرط مربوط به کلاس c بودن را محاسبه می کنیم. برای این کار، تعداد آن تکرار آن کلمه بر تعداد کل کلمات مربوط به آن کلاس تقسیم می کنیم. سپس برای احتمال اصلی (جمله به شرط کلاس) احتمال شرطی کلمه های تشکیل دهنده را در هم ضرب می کنیم. (چون کلمه ها را مستقل از هم در نظر میگیرم). دقت شود که برای احتمال کلمه (P(x_i | c را بدون توجه به ترتیب کلمه ها در جمله،از احتمال کلمه هایی که در بالاتر توضیح دادیم استفاده می کنیم.
prior : این احتمال، دید اولیه ما نسبت به کلاس جمله یا متن مورد بحث را نشان می دهد (قبل از مشاهده evidence). برای این احتمال فرکانس نسبی کلاس c را در کل داده های train محاسبه می کنیم.
postier : دیدگاه ثانویه ما به کلاس آن جمله یا متن مورد بحث بیان می کند. نحوه محاسبه نیز در صورت پروژه بیان شده!

برای سادگی و چون P(x) مستقل از کلاس است و برای همه کلاس ها یکی است و همچنین چون در نهایت فقط مقدار احتمال های postier مقایسه می شوند،از محاسبه P(x) در کد پروژه خودداری می کنیم.

3- مثال :

- علی بالاخره پروانه کسب خود را برای فروشگاه گرفت
- من دیروز پروانه زیبایی را دیدم.

واضح است که کلمه پروانه در دو جمله معنایی متفاوت دارد. با استفاده از bigram ها دو کلمه با هم در نظر گرفته می شوند و می توان ترکیب < پروانه کسب> و <پروانه زیبا> را از هم متمایز

تشخیص داد(به علت کسب و زیبا در دو ترکیب). در این مثال استفاده از bigam کافی به نظر میرسد اما استفاده از n-grams نیز هم می تواند ما را به قطعیت بهتری در فرق داشتن معنی پروانه در دو حمله برساند.(فروشگاه / دیدن)

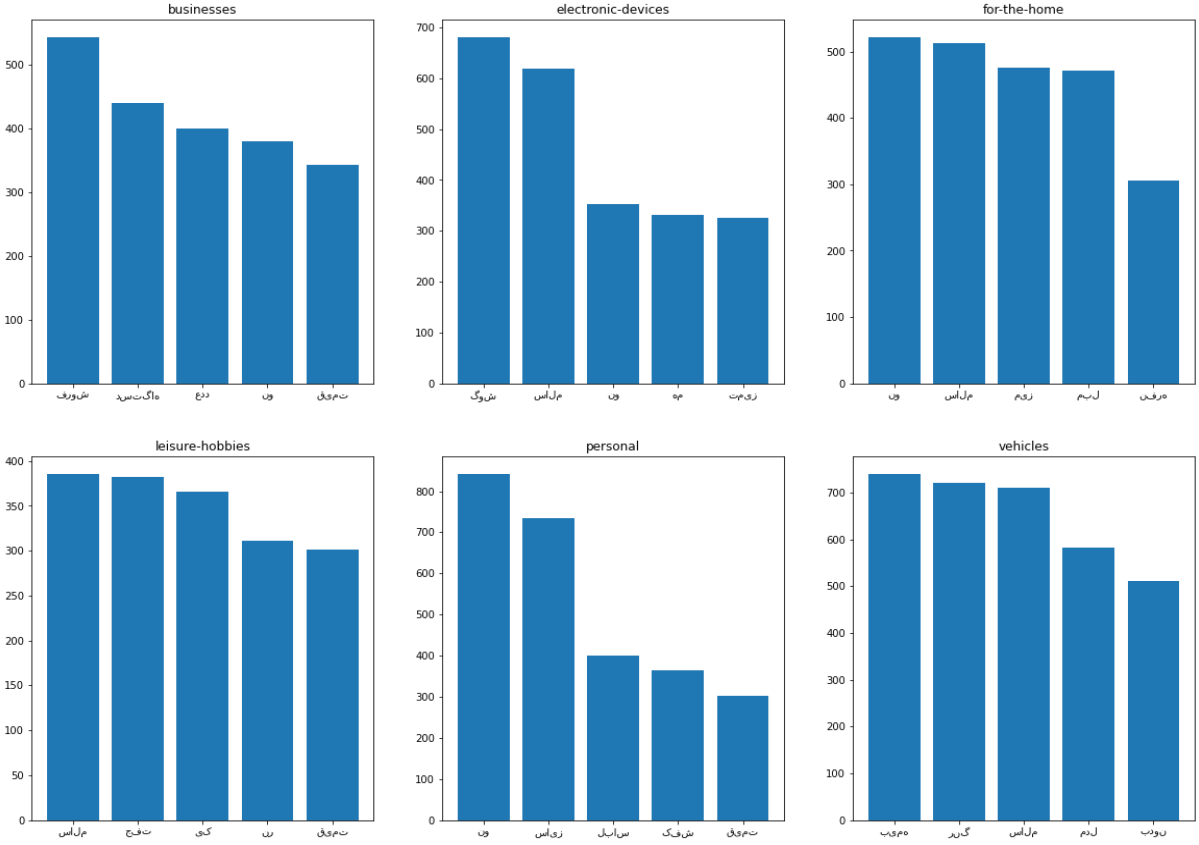
4- بدون استفاده از additive smoothing، در هنگام بررسی داده تست، به فرض اگر کلمه لاستیک در آگهی مربوط در کسب و کار به کار رفته باشد، ولی در داده های train چنین لغتی فقط در دامنه لغات مربوط به دسته وسایل نقلیه باشد،در نتیجه چون در مرحله training model ، احتمال (کلمه لاستیک به شرط دسته کسب و کار)صفر در نظر گرفته شده است، کل احتمال postier داده به شرط دسته کسب و کار را صفر می کند و ممکن است این احتمال به ازای دسته وسایل نقلیه صفر نشود در نتیجه classifier دسته را وسایل نقلیه تشخیص می دهد.
5- فرمول محاسبه احتمال شرطی کلمه به شرط دسته خاص در تصویر زیر آمده است:

$$\hat{P}(w_i \mid c) = \frac{count(w_i,c)+1}{\sum_{w \in V} \left(count(w,c)+1\right)} \quad => \quad P(w|c) = \frac{count(w,c)+1}{count(c)+|V|+1}$$

در این روش در دامنه لغات هر دسته(در داده های train)، علاوه کلماتی که در آن دسته وجود دادند،یک کلمه ای دیگر با عنوان unknown در نظر می گیریم که مقدار(count(w,c برای آن صفر است اما همچنان صورت کسر صفر نیست. در نتیجه اگر در جمله تست، کلمه ای وجود داشت که جزء دامنه لغات آن دسته نبود، احتمال شرطی آن به ازای آن دسته، برابر همین احتمال کلمه unknown می شود. در این صورت مشکل سوال قبل بر

طرف می شود. دقت شود که با تغییر فرمول، باز هم احتمال شرطی کلماتی که در دامنه لغات دسته خاص در داده های train بودند بیشتر است نسبت به کلمه هایی که نبودند(unknown)

6- نمودار :



7- از آنجا که precision نسبت حدس زده های درست به کل حدس زده ها، از یک نوع کلاس(کلاس A فرضا) را بیان می کند، ممکن است این نسبت زیاد باشد، ولی دسته بند، آنچنان دقیق نباشد. چرا که این نسبت اطلاعاتی درباره اینکه چقدر از داده های مربوط به کلاس A از داده های تست را، درست تشخیص داده، نمی دهد. مثلا در یک دسته بند binary، اگر در داده های تست ۱۰ داده با کلاس A داشته باشیم که فقط ۵ تا از آن ها A

تشخیص داده شده و در کل ۶ داده تشخیص داده شده با کلاس A داشته باشیم:

precision = ⅚ = 0.83 but recall = 5/10 = 0.5

از طرفی اطلاع داشتن از نسبت حدس زده های درست کلاس A به کل داده هایی که کلاس A داشتند (در داده های تست) نیز اطلاعاتی از نسبت precision چقدر بوده نمی دهد. برای مثال در یک دسته بند باینری:

TP = 5, TP + FP = 6, TP + TN = 10 => recall = 0.83 but precision = 0.5

بنابراین دانستن هر دو نسبت با هم لازم است و دسته بندی خوب است که هر دوی این نسبت در آن مقداری بالا داشته باشند.

8- F1 از میانگین harmonic استفاده می کند:

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots+\frac{1}{x_n}}$$

در این نوع میانگین گیری بر خلاف نوع عادی، به متغیری که مقدار کمتری دارد، وزن بیشتری داده می شود. این برای ما مهم است; چون بر اساس گفته های سوال قبل،باید هر دو مقدار precision و recall برای دسته بند مقدار بالایی داشته باشند. مثلا قرض کنید که دسته بند A و B مقادیر recall , precision زیر را داریم:

A : precision = recall = 0.8 | B : precision = 0.6 , recall = 1

حال با میانگیری معمولی قدرت A , B یکسان است اما با F1 قدرت A بیشتر است که منطقی است. چون هر دو شاخصه recall , precision برای آن بالا است.

9 - macro: این میانگین گیری،یک میانگیرین معمولی است. برای مثال شاخص f1 برای هر کلاس محاسبه شده و در نهایت از اعداد به دست آمده برای هر کلاس میانگین گیری می شود و به عنوان macro-f1 گزارش می شود

weighted : مانند حالت قبلی است با این تفاوت که اعداد به صورت وزن دار در میانگین گیری شرکت می کنند. وزن هر عدد تعداد داده با کلاس متناظر با آن عدد(مثلا عدد حاصل از f1 برای کلاس A) در داده های تست است.

micro: برای محاسبه micro-f1،ایتدا micro-recall و micro-precision را به دست می آوریم و سپس از این دو f1 میگیریم. محاسبه micro-recall و micro-precision همانند accuracy است.

10- ب)

*****without additive smoothing:

leisure-hobbies :

precision: 22.55

recall: 98.00

F1: 36.66

---

vehicles :

precision: 94.32

recall: 27.67

F1: 42.78

---

for-the-home :
precision: 86.05
recall: 37.00
F1: 51.75
---
personal :
precision: 91.59
recall: 32.67
F1: 48.16
---
electronic-devices :
precision: 88.52
recall: 18.00
F1: 29.92
---
businesses :
precision: 76.58
recall: 28.33
F1: 41.36

**all classes:
accuracy: 40.28
macro avg F1: 41.77
micro avg F1: 40.28
weighted avg F1: 41.77

الف)

*****with additive smoothing:

leisure-hobbies :

precision: 92.51

recall: 82.33

F1: 87.13

---

vehicles :

precision: 94.37

recall: 89.33

F1: 91.78

---

for-the-home :

precision: 76.55

recall: 90.33

F1: 82.87

---

personal :

precision: 87.06

recall: 89.67

F1: 88.34

---

electronic-devices :

precision: 91.84

recall: 90.00

F1: 90.91

---

businesses :
precision: 79.11
recall: 77.00
F1: 78.04

**all classes:
accuracy: 86.44
macro avg F1: 86.51
micro avg F1: 86.44
weighted avg F1: 86.51

11 - مقدار شاخص ها در حالت الف نسبت به ب افزایش پیدا کرده است. به خصوص شاخص های مربوط به کل کلاس ها (بالاتر از ۸۵ درصد)
12- جدول در قسمت خروجی کد نشان داده شده است. علت اشتباه می تواند این باشد که شاید در بخش پیش پردازش باید قوی تر عمل می کردیم. شاید هم استفاده از bigrams یا n-grams می توانست دسته بند را بهبود ببخشد.