گزارش تمرین ۳ - مدل discriminative برای کلاسه‌بندی توییت‌ها

برای استفاده از این متد ابتدا به معرفی فیچرها می‌پردازیم:

۱- فیچر bag of words

در این ویژگی ما همه‌ی کلماتی که در روش قبلی یعنی Naive Bayes داشتیم را نیز برای این قسمت استفاده می‌کنیم. به طوری که تغریبا همه‌ی کلمات یک توییت به عنوان یه فیچر در نظر گرفته می‌شوند. (البته با توجه به شرط مساله نمی‌توان از همه‌ی کلمات استفاده کرد)

*توجه شود در این روش تعداد زیادی استاپ‌ورد از هر توییت حذف می‌شود، بنابراین کلماتی که باقی می‌ماند واقعا معنی‌دار هستند.

۲- فیچر most effective bag of words

این ویژگی نیز به مانند قبل است با این تفاوت که به جای اینکه از تمام کلمات استفاده کنیم از زیر مجموعه‌ای از آن‌ها استفاده می‌کنیم. بنابراین تنها کلماتی از توییت که عضو این زیرمجموعه باشند به عنوان فیچر استفاده می‌شوند.

این زیر مجموعه شامل ۵۰۰۰۰ کلمه (یک دهم دایره لغات) است که با بیشترین اختلاف در کلاس ظاهر شدند که بیشترین تاثیر را در پیدا کردن تفاوت بین آن‌ها دارند. برای به دست‌آوردن آن از احتمالاتی که در Naive Bayes به دست‌آمده است استفاده می‌کنیم.

۳- فیچر hashtag

هر توییت ممکن است شامل چند هشتگ باشد، ما برای هر توییت حداکثر یک هشتگ را در نظر می‌گیریم. اگر توییت بیشتر از یک هشتگ داشته‌باشد، اولین آن‌را در نظر می‌گیریم. هشتگ‌ها هم تنها در صورتی انتخاب می‌شوند که عضو هشتگ‌های پر استفاده باشند.

هشتگ‌های پر استفاده، ۵۰۰۰ تا از هشتگ‌هایی هستند که در کل داده بیشترین حضور را داشته‌اند.

۴- فیچر منشن

اگر توییتی در جواب کسی باشد یا مستقیما نام‌کاربری کسی را اشاره کند، می‌گوییم این توییت یک فیچر منشن دارد و آن، نام کاربری کسی است که منشن شده‌است. اگر توییت شامل چند منشن باشد، اولین آن‌را انتخاب می‌کنیم و تنها در صورتی انتخاب می‌شود که عضو مجموعه‌ی ۱۰۰ منشن‌شده‌ی برتر باشد.

۵- فیچر emoji

ما در هر توییت ۲تا از بیشتر ایموجی‌های استفاده شده‌ی آن را به عنوان یک فیچر در نظر می‌گیریم.
*قبل از استخراج این ویژگی لازم است ابتدا تمیزکاری‌هایی روی دیتا انجام شود. به طور مثال ابتدا هر «:))))...» یا «:((((...» را به ایموجی مربوطه‌ی آن تبدیل می‌کنیم تا با بقیه ایموجی‌های استفاده شده هم‌خوانی داشته باشد.

نتایج

۱- فیچر bag of words

دسته	accuracy	recall	precision	f1
MJ_Akbarin	90.93	98.47	90.96	94.57
hamidrasaee	90.93	60.39	90.75	72.52

۲- فیچر hashtag + most effective bag of word

دسته	accuracy	recall	precision	f1
MJ_Akbarin	90.75	99.51	90.05	94.53
hamidrasaee	90.75	55.01	96.53	70.08

۳- تمامی فیچرها

دسته	accuracy	recall	precision	f1
MJ_Akbarin	86.05	97.31	87.02	91.8

دسته	accuracy	recall	precision	f1
hamidrasaee	86.05	37.85	76.71	50.69

در انتها می‌توان نتیجه گرفت بهترین فیچرها همان bag of word است. که البته تقریبا قابل توجیه نیز است. زیرا ما کلمات زاید را از آن حذف کردیم و هر کلمه‌ای که می‌ماند واقعا معنی‌دار است که خب نتیجه‌ی آزمایش هم این را تایید می‌کند.

ظاهر فیچرهایی مانند emoji و mention نمی‌تواند باعث بهبود نتایج شود و برعکس باعث گیج‌شدن کلسیفایر ما می‌شود.