به نام آنکه آموخت انسان را آنچه نمی دانست

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش زبان‌های طبیعی

CA1

گروه‌بندی اخبار

behzad.shayegh@ut.ac.ur

بهزاد شایق بروجنی

810 196 678

اسفند ۱۳۹۸

فهرست

مقدمه

با سلام. یکی از دغدغه‌های امروزه تعدد اخبار روزانه می‌باشد. هر روز برای هر نفر توده‌ای از اخبار کوتاه ارسال می‌شود که در اکثر مواقع فرد مذکور نسبت به اکثر آن‌ها بی‌تفاوت است و میلی به آن موضوعات ندارد و همین عمل سبب می‌شود شخص از خواندن اخبار امتناع کند و درنتیجه اخباری که لازم دارد را نیز از دست می‌دهد. خوب بود اگر سیستمی طراحی می‌شد تا این اخبار متعدد را به‌صورت خودکار به موضوعات مختلف دسته‌بندی کند و به هرشخص دقیقاً آنچه را می‌خواهد ارائه کند. هدف از انجام این پروژه ارائه‌ی چنین سیستمی با روش‌های ساده‌ی پردازش متن می‌باشد.

دادگان

دادگان فارسی از اخبار روزنامه همشهری بین سالهای ۱۳۷۵ تا ۱۳۸۶ با استفاده از خزشگرهای وب ایجاد شده است. با توجه به حجم بالای دادگان اصلی، نمونه‌ای از آن ایجاد و مورد استفاده قرار می‌گیرد. تعداد ۲۳۸۱ متن خبری در 6 کلاس تکنولوژی، ورزشی، اجتماعی، سیاسی، اقتصادی و فرهنگی در بخش دادگان آموزشی قرار گرفته است. هم‌چنین ۶۰۰ متن خبری بدون کلاس در بخش دادگان تست برای دسته‌بندی وجود دارد. برای ارزیابی مدل‌هایی که استفاده می‌کنیم، لازم داریم مجموعه‌ای از دادگان ارزیابی داشته باشیم پس تقریباً ۲۰٪ از دادگان آموزشی را بصورت تصادفی جدا کرده و به عنوان دادگان ارزیابی درنظر گرفتیم.

پیش‌پردازش

پیش‌پردازش دادگان از گام‌های مهم پروژه‌های پردازش زبان‌های طبیعی است. این پیش پردازش را به چند بخش تقسیم می‌کنیم :

Normalize

نوشته‌ها گونه‌های نوشتاری بسیار متفاوتی دارند. در ابتدا سعی کردیم عملیات «عادی سازی» را بر روی تمام دادگان خود (اعم از دادگان آموزشی، ارزیابی و تست) اعمال کنیم و متون جایگزین را استفاده کنیم.

Stem

لغات با ریشه‌های مشترک تأثیر معنایی مشابهی بر محتوای کلی متن دارند اما این اشتراک برای رایانه مشهود نیست. پس تمام لغات تمام دادگان را ریشه‌یابی کرده و از ریشه‌ی لغات بجای آن‌ها استفاده کردیم.

Lemma

لغات دارای پیشوندها، پسوندها، شناسه‌ها، عبارات جمع و ... هستند که این موارد بر روی تأثیر کلمه بر محتوای معنایی متن بی‌اثرند. پس این موارد را نیز در کل دادگان ساده‌سازی کردیم.

Vocabulary

حال متونی استاندارد و یک دست داریم. برای ادامه، در ابتدا نیاز به یک لغتنامه داریم. با توجه به این‌که در ادامه از دو نوع مدل زبانی (کلمه-پایه و حرف-پایه) استفاده خواهیم کرد، دو مجموعه‌ی لغتنامه و مجموعه حروف مورد استفاده را بصورت یکتا از مجموعه‌ی تمام دادگان استخراج کردیم. تعداد اعضای یکتای این دو مجموعه به ترتیب 35466 و 266 عضو بدست آمد. اگر مجموعه لغات را بدون پیش‌پردازش بدست می‌آوردیم، تعداد اعضای یکتای لغتنامه بسیار بیشتر از این می‌بود که این نشان‌دهنده‌ی حالات نوشتار بسیار متفاوت یک کلمه با یک مفهوم در زبان طبیعیست.

نشانه‌گذاری

این مورد که به بخش‌بندی متون (مشخص کردن ابتدا و انتهای جملات) مربوط می‌شود نیز گامی از پیش‌پردازش محسوب می‌شود که ما آن را حین ساخت مدل زبانی اعمال کردیم چرا که بازدهی زمانی بهتری داشت. البته بازدهی نهایی مدل‌ها را یک‌بار با انجام این عمل و یک‌بار بدون آن امتحان کردیم؛ نتیجه آن بود که بدون استفاده از این عمل بازدهی بهتری حاصل شد. جلوتر به‌صورت مفصل درمورد آن بحث خواهیم کرد.

مدل‌های زبانی

برای هدف پروژه، قصد داریم از یکی از چهار نوع مدل زبانی یکتایی کلمه، دوتایی کلمه، یکتایی حرف و یا دوتایی حرف استفاده کنیم. پس برای تک‌تک این انواع، تمام مراحل آموزش و ارزیابی را طی کردیم و مدل با بهترین نتیجه را انتخاب کردیم. هدف آن است که هر کلاس از مدل زبانی به هر خبر امتیازی نشان‌دهنده‌ی میزان احتمال تعلق آن خبر به آن کلاس بدهد (رابطه‌ی بین امتیاز و احتمال خطی نیست) و با مقایسه‌ی ۶ امتیاز حاصل برای هر خبر، کلاس آن خبر را پیشبینی کنیم. پس با هر نوع مدل زبانی، ۶ مدل بر روی دادگان آموزشی (برای هر کلاس خبری یک مدل) آموزش دادیم (۲۴ مدل زبانی آموزش دادیم).

معیار سرگشتگی

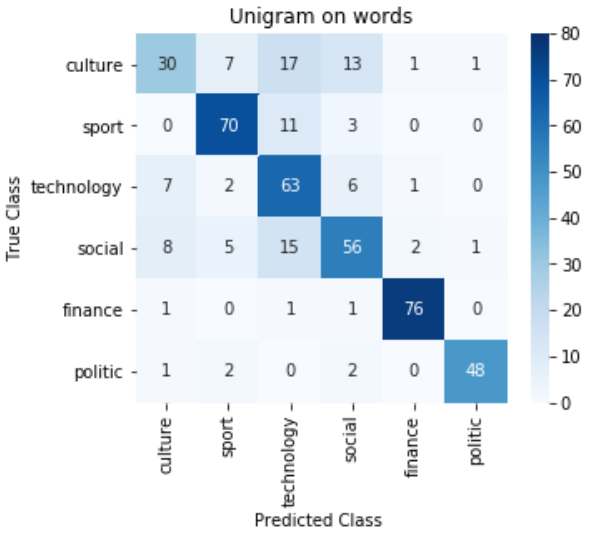
همانطور که در بخش قبل بیان شد، می‌خواهیم با هر کلاس مدل زبانی احتمال تعلق هر خبر را به آن دسته ارزیابی کنیم. برای این منظور از معیاری به نام معیار سرگشتگی استفاده می‌کنیم. معیار سرگشتگی برای یک خبر خاص، میزان کمیابی حضور واحدهای متنی (ngramها) در دادگان آموزشی آن کلاس را نشان می‌دهد، پس کم‌تر بودن این معیار به معنای احتمال بیشتر تعلق آن خبر به آن دسته می‌باشد.

Smoothing

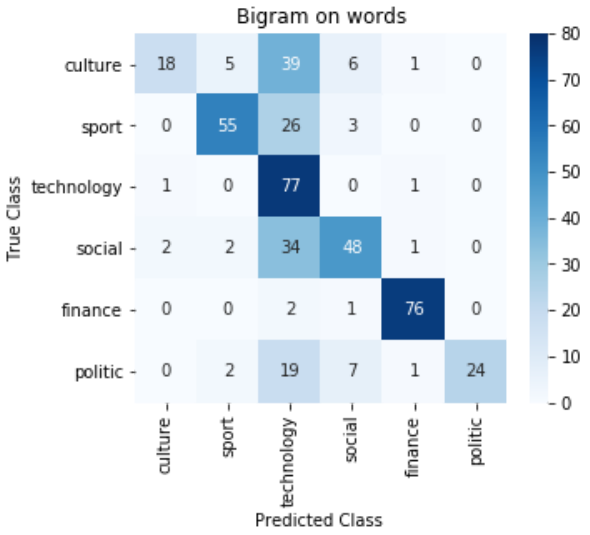
در فرایند محاسبه‌ی معیار سرگشتگی با مشکلی مواجه بودیم به این نحو که مقادیر اکثر سرگشتگی‌ها بی‌نهایت می‌شد. همانطور که بیان شد این معیار میزان کمیاب بودن را نشان می‌دهد پس این مشکل به معنای نایاب بودن حداقل یکی از ngram های داده‌ی ارزیابی می‌باشد. برای حل این مشکل از روش Laplace Smoothing استفاده کردیم و مشکل بصورت موفقیت آمیز حل شد.

Models : Unigram on words, Bigram on words, Unigram on letters, Bigram on letters

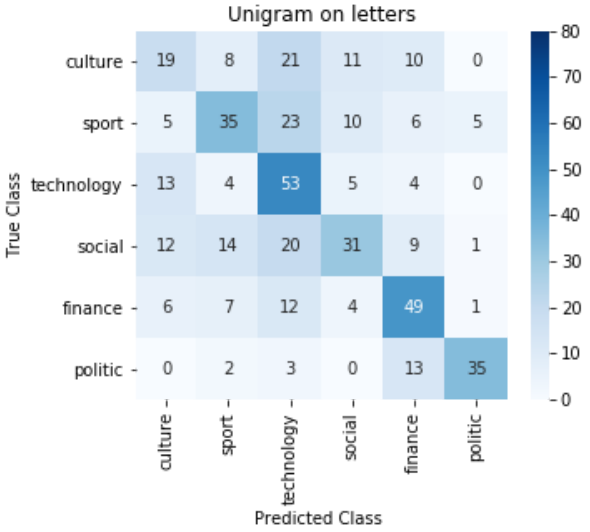
در این بخش، برای هر نوع مدل زبانی، به ازای هر خبر در دادگان ارزیابی ۶ معیار سرگشتگی متعلق به ۶ کلاس خبری را بدست آوردیم و سپس کلاسی را به عنوان دسته‌ی خبری خبر مذکور معرفی کردیم که کمترین مقدار سرگشتگی را مابین این ۶ مقدار داشته باشد. با مقایسه‌ی نتایج حاصل برای تمام دادگان ارزیابی با دسته‌های خبری واقعی آن‌ها به نتایج زیر رسیدیم (این مقایسه را با ماتریس درهم‌ریختگی مشهود می‌سازیم):



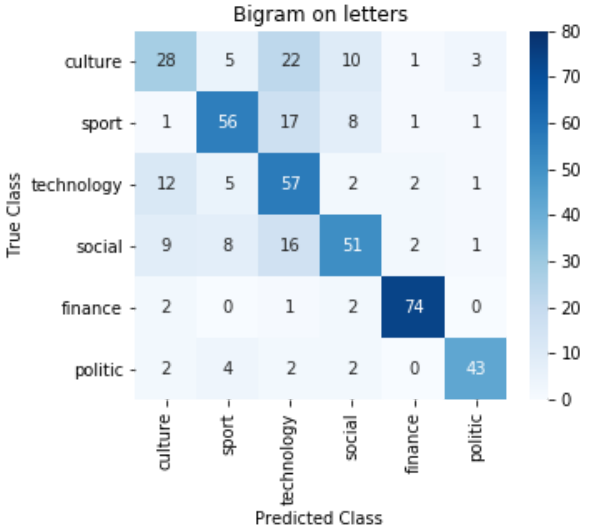
f1 : 0.7605321507760532
Precision : 0.7605321507760532
recall : 0.7605321507760532



f1 : 0.6607538802660754
Precision : 0.6607538802660754
recall : 0.6607538802660754



f1 : 0.49223946784922396
Precision : 0.49223946784922396
recall : 0.49223946784922396



f1 : 0.6851441241685144
Precision : 0.6851441241685144
recall : 0.6851441241685144

Precision, Recall, f1

زیر هر نمودار، سه معیار Precision, Recall و f1 آورده شده است. همانطور که مشاهده می‌کنید این سه معیار بسیار به هم نزدیک‌اند. این نتیجه را اینطور تعبیر کردیم که با توجه به سادگی معیار ارزیابی ما و کم بودن دادگان آموزشی و همچنین پراکندگی تقریبا یکسان دادگان تست بین کلاس‌ها و عنوان کردن میانگین نتیجه برای کلاس‌ها، میزان خطاگویی مدل ما بصورت متوازنی بین false positive و false negative تقسیم شده است. ما برای تصمیم گیری از معیار f1 که حاصل نوعی میانگین گیری از دو معیار دیگر می‌باشد استفاده می‌کنیم.

Letter-base vs. Word-base & Unigram vs. Bigram

همانطور که مشاهده می‌شود، بصورت کلی مدل‌های کلمه‌ای بهتر از مدل‌های حرفی عمل می‌کنند. این نتیجه برای ما عجیب نیست زیرا که ما در این مسئله سعی بر تشخیص محتوای خبرها داریم و این کلمات هستند که معنا و مفهوم داشته و حضورشان در جمله به آن محتوا می‌دهد نه حروف. یک حرف به تنهایی موضوعی را مشخص نمی‌کند اما حضور کلمه‌ی «سیاست» مشخصاً خبر را به سیاست مرتبط می‌سازد. همچنین مشاهده می‌کنیم که مدل یکتایی کلمه از مدل دوتایی آن بهتر عمل می‌کند که این موضوع کمی عجیب است، چرا که ما انتظار داریم زوج کلمات مفهوم بهتری را القا کنند. دلیل این تفاوت نتیجه با انتظار آن است که دادگان آموزش ما دادگانی طبیعی هستند و بازه‌ی وسیعی از لغات را دربر می‌گیرند اما تعداد نمونه‌های کمی برای آموزش داریم. پس تعداد زیادی از زوج کلمات موجود در لغت‌نامه را در دادگان آموزشی مشاهده نخواهیم کرد که ممکن است آن‌ها را در دادگان ارزیابی ببینیم. این موضوع با احتمال کمتری برای مدل یکتایی برقرار است و به همین خاطر مدل یکتایی در وضعیتی این چنینی که داده آموزشی کمی داریم پاسخ بهتری می‌دهد. همچنین در مدل‌های حرفی به این دلیل که با موضوع کمیابی ngramها و کمبود داده کمتر مواجهیم، مدل دوتایی حرفی ضعف مدل دوتایی کلمه‌ای را ندارد.

Size of perplexity values

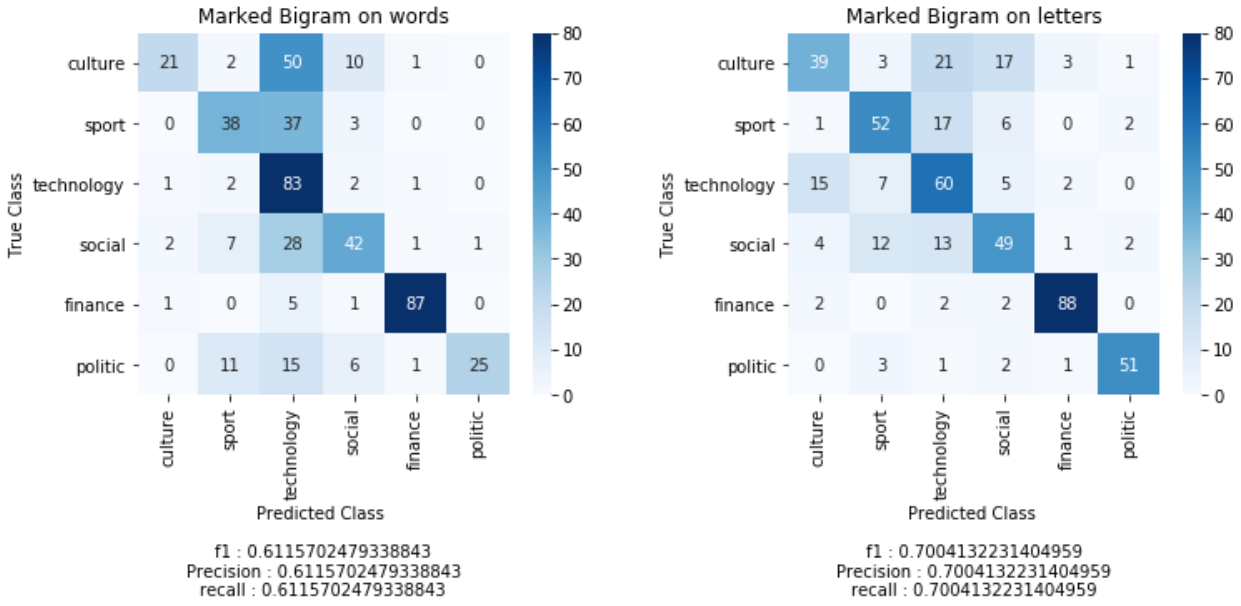
از نکات قابل توجه دیگر، مقایسه‌ی میزان بزرگی معیار سرگشتگی (بر روی تمام دادگان ارزیابی به ازای تمام کلاس‌های خبری) حاصل از انواع مدل‌های زبانی است. میانگین معیار سرگشتگی برای تمامی دادگان ارزیابی به ازای کلاس‌های مختلف مدل‌های زبانی مختلف به شکل زیر بدست آمد :

```
Unigram on words :
        culture         1577.431327
        sport           1758.345754
        technology      1936.803652
        social          1523.487591
        finance         1659.863651
        politic         1575.751914
Bigram on words :
        culture         12299.154049
        sport           12392.093804
        technology      15398.183232
        social          11371.261146
        finance         12173.525499
        politic         10510.570393
Unigram on letters :
        culture         22.835115
        sport           22.918374
        technology      23.272200
        social          22.771716
        finance         22.807691
        politic         23.148817
Bigram on letters :
        culture         13.499063
        sport           13.916161
        technology      14.214430
        social          13.409240
        finance         13.706349
        politic         13.459209
```

همانطور که مشخص است، این مقادیر برای مدل‌های حرفی بسیار کوچک‌تر هستند. این موضوع با تعریف ما از معیار همخوانی دارد چراکه حضور حروف در دادگان آموزشی خیلی معمول‌تر است و معیار کم‌یابی آن‌ها مقدار کوچکتری می‌گیرد. همچنین با توجه به اینکه احتمال مشاهده‌ی دوتایی کلمات بسیار پایین‌تر است (تنوع بیشتری دارند) پس میزان کمیابی اعضای آن‌ها و درنتیجه معیار سرگشتگی آن‌ها بسیار بزرگ‌تر خواهد بود.

Marked Bigram Models

همانطور که در بخش «پیش‌پردازش» اشاره شد، ما نوع دیگری از ngramسازی را نیز تجربه کردیم. نتایج بالا برای مدل‌های زبانی دوتایی به این نحو آموزش دیده و ارزیابی شدند که در مدل کلمه‌ای، لغت آخر هر جمله به عنوان لغت اول هیچ زوجی استفاده نشد؛ همچنین لغت اول هر جمله به عنوان لغت دوم هیچ زوجی استفاده نشد؛ برای دوتایی حرف نیز حرف اول و آخر هر کلمه به عنوان حروف دوم و اول هیچ زوج حرفی درنظر گرفته نشد. ما بار دیگر برای مدل کلمه‌ای با درنظر گرفتن علامتی برای ابتدای جملات و علامتی برای انتهای جملات و همچنین برای مدل حرفی با درنظر گرفتن علامتی برای ابتدا و انتهای لغات این موارد را نیز اعمال کردیم و نتایج به شکل زیر بود:

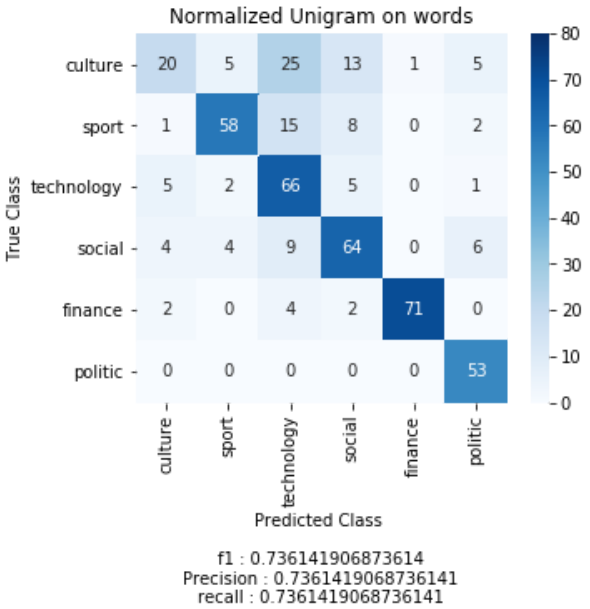


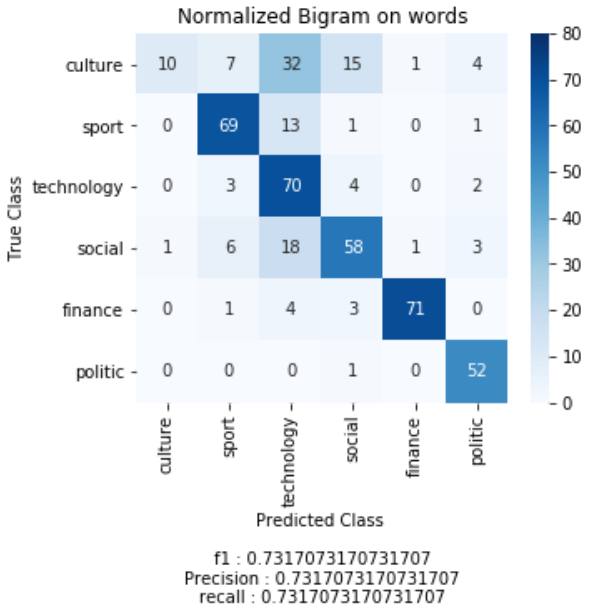
همانطور که مشاهده می‌کنید این موضوع برای حروف تأثیر مثبت و برای کلمات تأثیر منفی داشت. تأثیر مثبت برروی حروف شاید به این خاطر باشد که کلمات آغازشونده با حرفی خاص و پایان پذیر با حرفی خاص محدودتر بوده و این موضوع به مدل در تشخیص کلمات کمک می‌کند. از این پس به این نوع مدل زبانی مدل «نشانه‌گذاری شده» می‌گوییم.

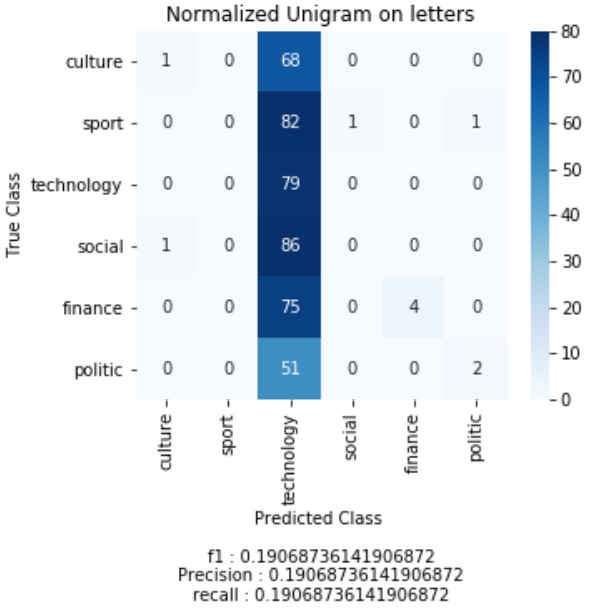
Normalized Models

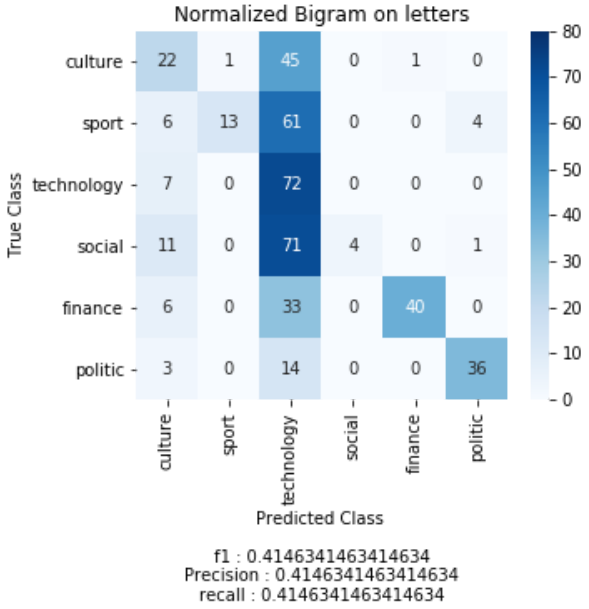
با توجه به اینکه از دادگان متفاوتی برای آموزش هر کلاس مدل زبانی استفاده کردیم، ممکن است تفاوت حجم در دادگان آموزشی باعث تجربه بیشتر یک مدل نسبت به دیگری شود از این نظر که تنوع ngramهای بیشتری را مشاهده کرده باشد(نه به

دلیل ذات متنوع خبر بلکه به دلیل حجم بیشتر اخبار) . پس شاید مفید باشد اگر بر روی مقادیر سرگشتگی خروجی هر کلاس مدل زبانی بر روی تمام دادگان تست عملیات «عادی سازی» انجام دهیم تا این موضوع را تا حدی کم اثرتر کنیم. با انجام این عمل به نتایج زیر رسیدیم :







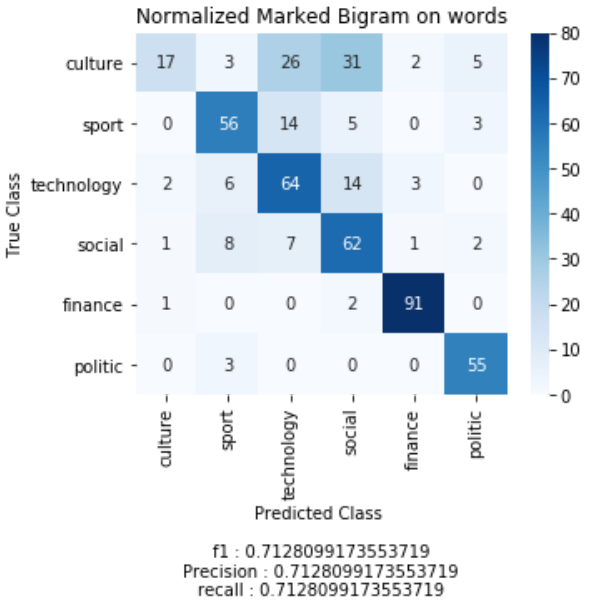


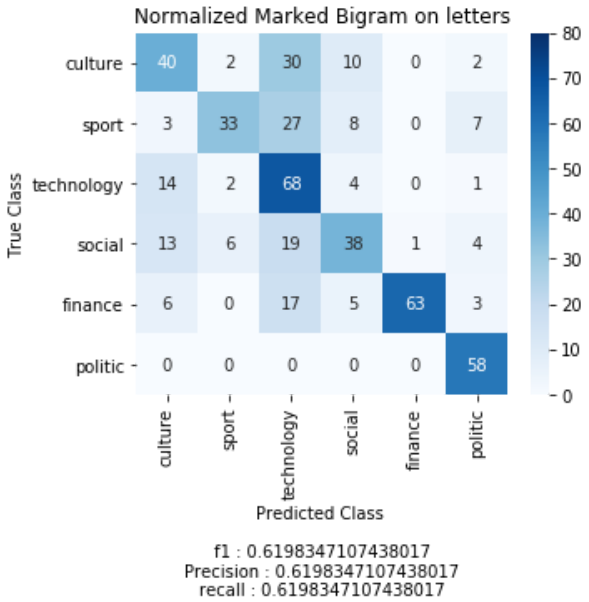
همانطور که مشاهده می‌شود این روش بر روی مدل‌های حرفی بسیار تاثیر بدی گذاشته است. این رویداد به این دلیل است که استدلال این عمل ما محروم ماندن یک مدل از مشاهده‌ی یک ngram بود اما با توجه به تعداد کم حروف این موضوع در مورد مدل‌های حرفی صادق نیست و این کار ما قدرت مدل را که بر اساس میزان تعدد حروف است از بین می‌برد. اما مشاهده می‌کنیم که این روش برای مدل‌های کلمه‌ای نیز مفید واقع نشده و دلیل این امر شاید توزیع حجمی عادلانه‌ی دادگان بین کلاس‌های خبریست.

احتمالا اگر از یک دسته‌ی خبری تعداد خبر کمتری داشتیم(نه بخاطر کمیاب بودن آن خبرها بلکه به دلیل ضعف در جمع‌آوری دادگان) این روش مفید واقع می‌شد.

Normalized Marked Models

گرچه در بخش قبل مشاهده کردیم روش «عادی سازی» اثر خوبی ندارد، بد نیست این موضوع را بر روی مدلهای «نشانه‌گذاری شده» نیز امتحان کنیم :





به صورت عجیبی بر روی این مدلها تاثیر بهتری داشت که نتوانستیم آنرا توجیه کنیم.

دادگان تست

همانطور که از نتایج قسمت قبل واضح است، بهترین مدل برای تشخیص کلاس خبری، مدل یکتایی کلمه بدون «نشانه گذاری» و «عادی سازی» می‌باشد. پس با این مدل مجموعه دادگان تست را ارزیابی کرده و کلاس پیشبینی شده برای هر مورد را ثبت می‌کنیم.

فایل‌های جانبی

به همراه این گزارش، یک پوشه به نام Codes ارائه می‌شود که حاوی فایل‌های زیر است :

Result.csv : فایل خروجی بخش آخر پروژه (دادگان تست به همراه کلاس پیشبینی شده) می‌باشد.

NLP_CA1.ipybn : فایل ژوپیتر کد پروژه که شامل تمام مراحل بجز آزمون‌های «نشانه‌گذاری» می‌باشد.

NLP_CA1_MarkedMode : بازسازی شده‌ی کد قبل اینبار با اعمال عملیات «نشانه‌گذاری» و مشاهده‌ی خروجی آزمون. تفاوت منطقی این دو قطعه کد تنها در یک خط می‌باشد که با کامنت «What is different» مشخص شده است.

NLP_CA1.html, NLP_CA1_MarkedMode : خروجی قابل نمایش فایل‌های قبل