به نام خدا

N-Gram Based Text Classification for Persian Newspaper Corpus

پاراگراف اول – هدف مقاله

این مقاله قصد دارد تا یک مدل طبقه‌بندی مبتنی بر n-Gram را در سطوح کلمه(نه جمله) بر روی کلمات فارسی برای طبقه بندی کلمات فارسی پیاده‌سازی کند. این مقاله به دنبال این است تا به کمک مدل احتمالی n-Gram بتواند مدلی را پیاده سازی کند که این مدل توانایی تشخیص معنای کلمات را داشته باشد تا به کمک این معنایی که دارند کلاس کلمات را تشخیص دهند تا نهایتا به مدلی برسیم که بتواند اخبار مجله ها را تجزیه و تحلیل کند.

پاراگراف دوم – خلاصه کار انجام شده در مقاله

دسته بندی کردن داده ها یکی از متد هایی می‌باشد که به کمک آن میتوانند یکسری استنتاج ها و یا پیشبینی هایی را انجام دهند. به این صورت که در این متد داده ها را در دسته های مختلف که هر دسته از یک نوع و مفهوم می‌باشد دسته بندی می کنند و برای هر دسته یک نام که در اینجا label آن دسته می‌باشد را در نظر می گیرند.

حالا Text classification نوعی متد می‌باشد که بر روی داده متنی انجام می‌شود که در اینجا داده های ما میتواند لغات و حروف های یک کلمه باشند که در کنار هم یک متن، پاراگراف و یا یک Document را تشکیل داده‌اند و دسته بندی ما میتواند به این صورت باشد که این Document ها را بررسی کنیم و به عنوان داده اصلی خودمون مشخص کنیم و یک نام به عنوان label برای آن در نظر بگیریم که نهایتا ما بتوانیم به عنوان ورودی مثلا یک Document را از ورودی دریافت کنیم و در خروجی دسته بندی آن را اعلام کنیم به این صورت که مثلا آن داده ما در دسته بندی هنر و معماری قرار میگیرد.

در این مقاله سعی شده است تا از متد Text classification برای داده های زبان فارسی استفاده شود، زیرا همانطور که بررسی این متد ها در زبان ها غربی بسیار مهم و رایج است، در آسیا نیز اهمیت دارد و به این موضوعات پرداخته شود. اما پیاده سازی این روش ها بر روی زبان فارسی در کنار مسئله اصلی که چگونه پیاده سازی کردن است دارای مشکلات دیگری می‌باشد، مثلا یکی از مشکلات ما این است که در زبان فارسی ما به سادگی نمیتوانیم ریشه یک کلمه را بدست آوریم و یک Benchmark برای زبان فارسی نداریم، که این مشکلات اضافه ای می‌باشد که با آن مواجه هستیم.

در Classification روش و الگوریتم های زیادی وجود دارد. وجه مشترک تمامی این روش ها این می‌باشد که این روش ها که با مسائل Text classification همانند یک مسئله رایج Classification مواجه میشوند. که در نتیجه این رویارویی تنها Text Classification را در دو مرحله اصلی Feature engineering و Classification learning انجام می‌دهند. اما باید این را دانست که روش های سنتی Classification برای Text classification مناسب نیستند به دلایل زیادی از جمله زیاد و بزرگ بودن

حجم داده هایی که در Text classification با آن مواجه هستیم، عدم توانایی در انتخاب درست و کامل feature ها در مرحله Feature selection و دلایل دیگر که نهایتا باعث می‌شود حل یک مسئله Text classification به روش استاندارد و رایج ما را از NLP بودن نتیجه مان دور کند زیرا که در حل این دست از مسائل به روش های استاندارد ما هیچ یک از ارتباطی که کلمات میتوانند با هم داشته باشند را در نظر نگرفته ایم و این سبب بی روح شدن مدل که ایجاد شد می‌شود.

در نتایج تمامی این مسائلی که وجود دارد ما برای دسته بندی متون فارسی به سراغ روش مدل زبانی احتمالی N-gram رفتیم. در پیاده سازی این روش برای متون فارسی ما با سوالاتی مواجه شدیم مثلا اینکه مدل N-gram را در سطح کاراکتر پیاده سازی کنیم یا در سطح کلمه ؟ و سوال دیگر این است که ارتباط بیت هر یک از این کلمات و کاراکتر ها را چگونه تعریف کنیم.

در این مدل ما از dataset همشهری استفاده کرده ایم و در فاز preprocessing که جز اولین فاز ها میباشد که همیشه قبل از دسته بندی کردن داده برای حذف و از بین بردن داده های غیر کارا بکار میرود، علامت های حروف فارسی،.. را از متن حذف کردیم. در این مقاله هم ما 4 مقدار 1 تا 4 را برای N در نظر گرفتیم و آن را پیاده سازی کرده ایم.

ما باید به دنبال ایجاد و پیاده سازی یک مدل زبانی ای باشیم که دسته بندی متن را برای ما انجام دهد، هدف از مدل های زبانی این است که احتمال رخ دادن یک دنباله را بدست می آورد و ما نهایتا میتوانیم با مجموعه دنباله هایی که داریم، که هر یک از این دنباله ها دارای احتمال های خاص خود هستند، دنباله ای که دارای بیشترین احتمال شده است را انتخاب کنیم.

همانطور که پیش تر ذکر شد مدل زبانی ای که ما انتخاب کردیم در این مسئله مدل N-gram میباشد. این مدل در عین ساده بودنش به کمک قانون Chain rule میتواند هر دنباله و احتمال انتخاب شدن آن را پیش بینی کند. این مدل احتمال هر دنباله ای را با توجه به N-1 کلمه قبلی خودش بدست می‌آورد.

حال ما باید از این مدل زبانی به عنوان Text classifiers استفاده کنیم تا Category یک داده ورودی(Document و ..) را تشخیص دهیم. قانون استفاده از مدل زبانی N-gram به عنوان یک Text-classifier درست، این است که، ما برای هر Categoryای که داریم، یک مدل زبانی مخصوص به آن را تعریف کنیم و نهایتا Document تولید شده توسط هر دو مدل را با هم بررسی کنیم تا از صحت درستی احتمال بیشتری حاصل کنیم.

عبارت Smoothing مطرح شده به تکنیکی اشاره میکند که در آن هدف ما افزایش تخمین احتمال داده شده است، به این صورت که احتمال کم را بیشتر و احتمال زیاد را کمتر میکند. انجام دادن اینکار سبب میشود که احتمال های صفری که داریم (احتمال های صفر زمانی بوجود می‌آیند که در آن کلمات با هم دیگر هیچ ارتباطی ندارند.)از بین بروند. چندین متد وجود دارد برای Smoothing : 1. روش Add-One که در این روش به تعداد Counter یک واحد اضافه می‌شود که نهایتا باعث میشود که ما احتمال صفر نداشته باشیم. این روش با افزایش تعداد دفعات رخ دادن n-gram ایجاد می‌شود. 2. روش Absolute Discounting که ای روش یک مقدار ثابت را از Counter ما کم میکند، که اینکار منجر میشود که احتمال ها کاهش یابند و 3. روش Back-0ff که این روش میگوید که در صورت نیاز به N-gram های پایین تر برویم یعنی اینکه اگر مثلا داریم 3-gram را اجرا میکنیم و نتوانستیم

هیچ سه عبارت مشابه را پیدا کنیم که احتمال آن را بررسی کنیم، در این زمان به N-gram پایینتر برویم یعنی 2-gram را بررسی کنیم و به همین ترتیب پیش برویم.

برای انجام Text classification در زبان فارسی ما باید یکسری کار ها را در گام پیش پردازش بر روی دیتاستمان انجام دهیم. از جمله اینکار ها:

- Text segmentation

در این کار ما بخش های Documentمان را مشخص میکنیم. یعنی مشخص کردن، پاراگراف ها، جملات و کلمات

- Word segmentation

معروف به Tokenization است که در این بخش کلمات، Punctuationها، طرز نوشته شدن حروف الفبا و پیشوند و پسوند کلمات را تشخیص میدهیم. و سپس تمامی این موارد را از Document حذف میکنیم.

- Normalizing

در این بخش ما به حروف هایی میپردازیم که در زبان فارسی چند نمونه از آن را داریم، مانند «ی» و یا کلماتی که به چند شکل نوشته میشوند ما این حروف و کلمات را ابتدا تشخیص و سپس یکی(برابر) می کنیم.

- Word stemming

و نهایتا کلماتی که دارای پیشوند و یا پسوند هستند را شناسایی میکنیم و پیشوند و پسوند آنها را از آنها حذف می کنیم.

ما تمامی این متد و روش و بخش های مختلفی که پیش تر ذکر شد را بر روی دیتاست همشهری که توسط دانشگاه تهران داده های آن جمع آوری شد پیاده سازی کردیم. و توانستیم مدل های خوبی را ایجاد کنیم که کارایی آن مناسب و قابل قبول است. همچنین وقتی آزمایش را بدون و با انجام دادن Preprocessing زبانی بر روی داده انجام دادیم، مشاهده کردیم که نتایج حاصل از زمانی که Preprocessing زبانی انجام میدهیم بهتر و کارا از زمانی است که این پیش پردازش را انجام نمی‌دهیم.

پاراگراف سوم – نتیجه‌گیری

از نتایجی که حاصل شد توانستیم دریابیم که انجام preprocessingهای زبانی میتواند از جمله کار های مفیدی باشد که در Text classification انجام میدهیم. همچنین با بررسی های شده دریافتیم که پایین بودن تعداد Gram سبب میشود که نتوانیم مدلمان را به خوبی آموزش دهیم زیرا در Gram های پایینتر ما ارتباط کلمات با هم را کمتر در نظر گرفتیم و در آموزش از این ارتباط ها استفاده نکردیم. و همچنین زیاد بودن تعداد Gram نیز سبب میشود که ما به احتمال های صفر برسیم که در نتیجه آن تحلیل و آموزش آن سخت و نادرست خواهد بود. در نتیجه دریافتیم انتخاب درست تعداد Gram نیز میتواند بسیار مفید و مهم باشد. البته باید بیان شود که در جاهایی که حتی با Gram مناسب به احتمال های صفر یا شدیدا متغیر رسیدیم، سعی کردیم از روش هایی که در Smoothing ذکر شد استفاده کنیم تا کار Normalization داده را درست و کامل انجام دهیم که بتوانیم به نتایج درست تری برسیم.

با توجه به تمامی نکات ذکر شده در این مقاله به این نتیجه رسیدیم که 3-Gram بسیار تعداد مناسبی برای ما بود و ما را به نتایج خوبی رساند.

سید علی شهداالحسینی

Seyyed Ali Shohadaalhosseini