به نام خدا

N-Gram Based Text Classification for Persian Newspaper Corpus

پاراگراف اول – هدف مقاله

ابتدا مسئله ای شکل میگیرد و سپس به دنبال راه حل آن میروند. افزایش استفاده از اینترنت منجر به افزایش تولید داده های متنی شد، تا جایی که به دنبال کلمات جدیدی گشتند برای توصیف این حجم زیاد از داده و نام کلان داده را بر روی آن قرار دادند، اگرچه که امروزه این حجم داده ها را نمیتوان دیگر با واژگانی چون کلان، توصیف کرد.

این حجم زیاد از داده های متنی منجر شد تا انسان ها دیگر نتوانند به سادگی مطالب را بررسی کنند و نتیجه گیری داشته باشند، و اینجا بود که مسئله و نیاز مطرح شد. انسان ها برای پاسخ دادن به چنین مسائلی به دنبال سیستم های اوتوماسیون رفتن تا این قبیل کار ها را به آن ها بسپارند.

در این مقاله میخواهیم بر روی یک وجه از این سیستم که NLP باشد کار کنیم و آن را در زبان فارسی مورد بررسی قرار دهیم و به یکسری تعاریف اولیه مفاهیم موجود در اینکار آشنا شویم. ما قصد داریم در این مقاله بر روی تجزیه و تحلیل احساسات (Sentiment analysis) کار کنیم. ما قصد داریم تا این کار را در سه مرحله پیاده سازی کنیم:

1. پیش پردازش
2. استخراج ویژگی
3. و طبقه بندی داده ها بر اساس متد Support vector machine (SVM)

پاراگراف دوم – خلاصه کار انجام شده در مقاله

تحلیل احساسات فرآیندی میباشد که در آن داده های بی قاعده و ساختار را کلاس بندی می‌کنند. به دلیل کمبود منابع برای تحلیل کردن ساختاری پیاده سازی شد تا ویژگی ها مورد نظر ما را از نظرات کاربران بیرون بکشد و سپس ما با کمک Support Vector machine عملکرد ویژگی های انتخاب شده را مورد بررسی قرار دادیم.

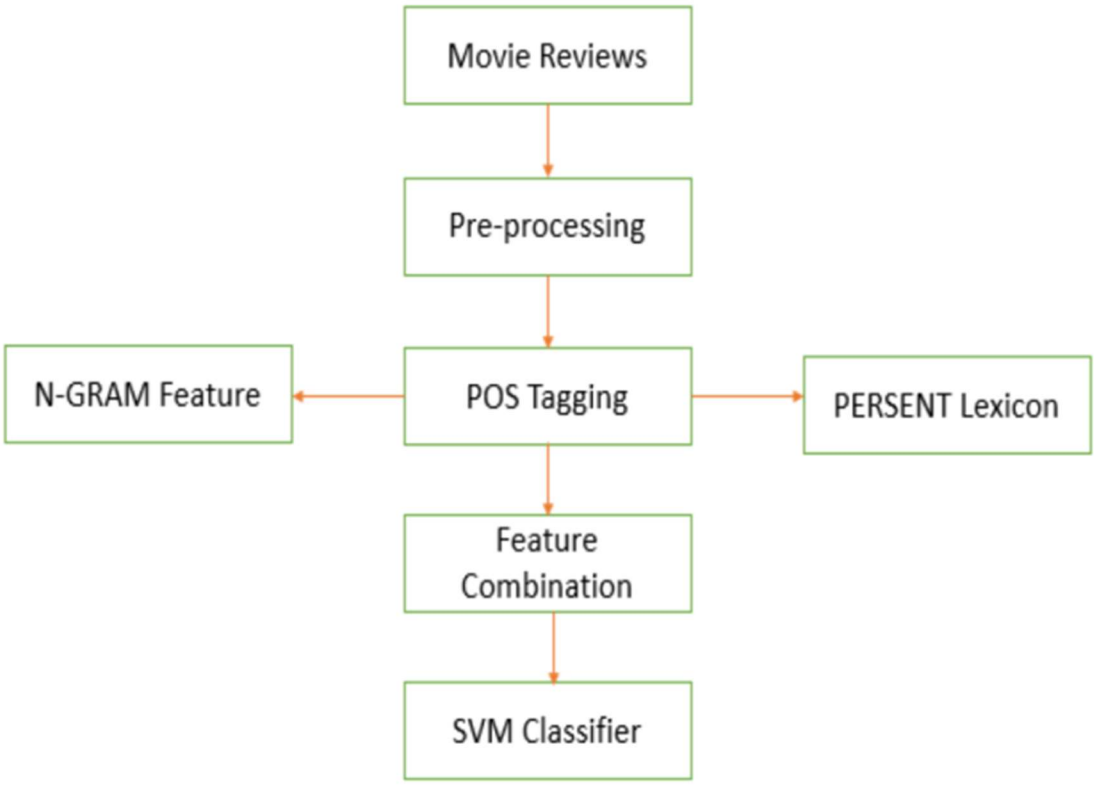
تحلیل احساسات به کمک 4 رویکرد مورد بررسی قرار میگیرد:

1. Keyword spotting: این رویکرد به دنبال این است که دسته بندی کلمات را در رابطه با اون حسی که دارند مشخص کنند. از حس های کلمات میتوان به sad و happy و afraid و ... اشاره کرد.
2. Lexical affinity: این رویکرد به کلمات دلخواه خودش یک حس معنای دقیق تری را سعی میکند نسبت دهد.
3. Statistical method: این متد به یادگیری ماشین کمک میکند تا احساسات را بتواند بهتر تشخیص دهد.
4. Concept based approach: این رویکرد نیز بر روی تحلیل معنایی جملات تمرکز میکند.

سید علی شهداالحسینی

Seyyed Ali Shohadaalhosseini

در ادامه در تصویر پایین میتواند مراحل کاری که در این مقاله انجام شد را مشاهده بفرمایید:



بعد از دریافت اطلاعات و متون:

1. Pre-processing: در این مرحله اطلاعات و دیتا ها جمع آوری و سپس نویز آن ها گرفته شد. همچنین اگر کلمات دارای حرف های تکراری با هدف افزایش نفوذ احساست هستند، این حروف ها را حذف کردیم. همچنین یکسان سازی کلمات نیز در این مرحله انجام شد یعنی برای حروف های چون «س» که چند حروف داریم، همه را حذف کردیم و از یک حرف واحد استفاده کردیم برای آن.
2. POS feature: در این مرحله ساختار گرامری جملات را مورد بررسی قرار دادیم، و این ویژگی ها (اسم، فعل، فاعل و....) را در داده ها تشخیص دادیم.
3. N-gram: سپس در این مرحله متد ها Unigram و Bigram و Trigram را بر روی دیتا ها پیاده سازی کردیم. (در مقاله قبلی در مورد N-gram بحث شد.)
4. PerSent lexicon: PerSent یک واژه نامه فارسی است که از 1500 کلمه که گرامر و میزان polarity آن مشخص شده است تشکیل شده است. به کمک این دیتاست ما داده های استخراج شده خودمان را مورد بررسی قرار دادیم و قطبیت و گرامر آن را مشخص کردیم.

پاراگراف سوم – نتیجه‌گیری

نتایج بدست آمده حاکی از آن است که بهترین کارایی را میتوانیم با ترکیب Unigram و Bigram و Trigram در انتخاب ویژگی بدست آوریم. مابین این متد نتایج نشان دهنده این هستند که متد trigram میتواند ما را به نتیجه بهتر و accuracy بالاتر برساند. اما با بررسی کردن 3 متد n-gram مشاهده شد که استفاده از هر سه متد با هم میتواند ما را به بهترین کارایی برساند. در آینده نیز میتوان تحقیقات را در مباحث deep learning مورد بررسی قرار داد.