تحلیل نظرات کاربران با اسپارک

دانشجو: ملیکا عیوقی

استاد راهنما: دکتر آزاده شاکری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

تصمیم های افراد در مورد موضوعات مختلف معمولا متاثر از نظرات دیگران درباره آن موضوع است. نمونه این رفتار را در خرید کالا ها مشاهده میکنیم، برای مثال وقتی قصد خرید کالایی را داریم، تمایل داریم از نظر سایر افرادی که از آن کالا استفاده کرده اند مطلع شویم. تحلیل نظر کاربرها به ارائه دهندگان یک سرویس کمک میکند از میزان رضایت مشتری خود با خبر شوند و از آن در جهت بهبود سرویس خود استفاده کنند و از طرف دیگر به مصرف کنندگان سرویس یک تجربه نزدیک به واقعیت ارائه میکند.

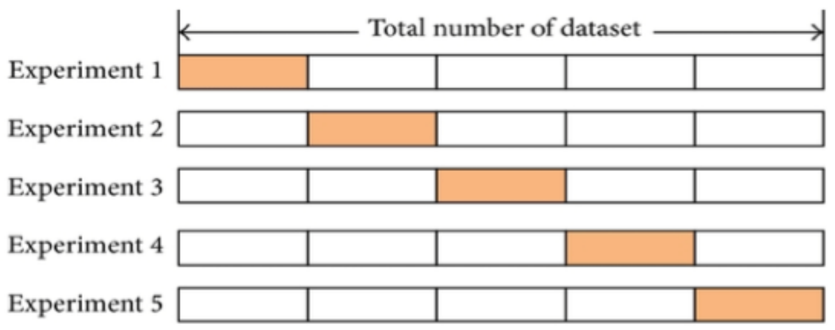
هدف از این پروژه، تعیین مثبت یا منفی بودن نظر ها با کمک الگوریتم های رده بندی نظارت شده می باشد. به دلیل حجم بالای نظرات و به تبع پردازش بالا، نیاز به استفاده از پلتفرمی برای داده های بزرگ می باشد که در این طرح از آپاچی اسپارک استفاده می شود.

مدل پیشنهادی

در شکل زیر روند تغییرات بر روی داده ها قابل مشاهده می باشد.

پس از اینکه داده های خام مراحل بالا را طی میکنند، نیاز به ارزیابی مدل پیشنهادی وجود دارد. مدل با معیار های ارزیابی صحت، دقت، بازخوانی، f-measure و سطح زیر نمودار ROC بررسی می شود.

برای اینکه نتایج آزمایش ها واریانس کمتری داشته باشند، از روش ارزیابی k-fold استفاده کردیم که توضیح این روش در شکل زیر قابل مشاهده است.



ارزیابی نتایج

در این بخش به ارزیابی نتایج بدست آمده از آزمایش ها می پردازیم. با مقایسه معیار ها برای کلاس مثبت و منفی، متوجه میشویم که معیار های کلاس منفی به اندازه قابل توجهی کمتر از کلاس مثبت هستند. همانطور که قابل مشاهده است، در کار های آینده باید کاری کرد که پیش بینی ها در کلاس منفی بهتر شوند. دلایل کم بودن معیار ها برای منفی ها:

- حذف فعل های منفی به عنوان ایست واژه
- عدم وجود کلمات منفی ساز مستقل
- کنایه آمیز بودن درصد قابل توجهی از نظرات در زبان فارسی

در تمامی معیار های ارزیابی، رگرسیون لاجیستیک بدتر یا برابر با دو روش دیگر عمل کرده است. در حالیکه مقایسه بین دو روش SVM و بیز ساده ممکن نیست. چرا که معیار ها خیلی نزدیک به هم هستند.

	SVM	Logistic Regression	Naïve Bayes
Precision(Neg)	۰.۳۵	۰.۲	۰.۳۷
Recall(Neg)	۰.۶۵	۰.۶۵	۰.۵۳
F1(Neg)	۰.۴۵	۰.۳۱	۰.۴۴
Precision(Pos)	۰.۹۷	۰.۹۶	۰.۹۶
Recall(Pos)	۰.۹۰	۰.۸	۰.۹۳
F1(Pos)	۰.۹۳	۰.۸۷	۰.۹۴
ROC-AUC	۰.۷۷	۰.۷۲	۰.۷۳
Accuracy	۰.۸۸	۰.۷۸	۰.۸۹
F1(Total)	۰.۹	۰.۸۳	۰.۹۱

جمع بندی

از دستاورد های این تحقیق ایجاد یک سیستم تحلیل خودکار نظرات فارسی کاربران جهت تسهیل فرآیند تصمیم گیری برای خرید کالا توسط مشتریان بالقوه می باشد.

از بزرگ ترین چالش های این پروژه، تحلیل نظرات به زبان فارسی بود چرا که منابع محدودی برای زبان فارسی وجود دارد. از دیگر چالش ها، نظرات کنایه آمیز کاربران بود.

برای بهبود ارزیابی ها برای کلاس منفی، نیاز به روش های هوشمندانه تری داریم مانند تعبیه روشی که منفی ساز های فارسی را به درستی تشخیص دهد. علاوه بر این می توان از روش های دیگری که روابط معنایی را دخیل می کنند کمک گرفت مانند روش بردار های تعبیه کلمات.

کاربرد های صنعتی:

شرکت های فروش کالا همانند دیجی کالا با کمک این سیستم تحلیل نظرات می توانند علاقه مندی های خریداران خود را تشخیص دهند و در نهایت سیستم پیشنهاد دهنده ای تعبیه کنند که به هر کاربر، کالای با کیفیتی که نیاز دارد را پیشنهاد دهد. و از طرف دیگر به دیجی کالا، پیشنهاد حذف و بهبود کالاهایی که را که مورد علاقه کاربران نیست بدهد.

مراجع اصلی

۱. ا. طبی فخر، "نظر کاوی در اسناد غیر رسمی و کوتاه" پایان نامه کارشناسی ارشد، دانشکده های فنی، دانشکده مهندسی برق و کامپیوتر، ۱۳۹۶

۲. ا. دهدار بهبهانی، "تحلیل خودکار قطبیت اسناد در مستندات فارسی" پایان نامه کارشناسی ارشد، دانشکده های فنی، دانشکده مهندسی برق و کامپیوتر، ۱۳۹۳

۳. م. غفوری، س. راحتی، م. پهلوان نژاد، ع. عظیمی زاده، "نرمال ساز متون فارسی" هجدهمین کنفرانس مهندسی برق ایران، ۲۰۱۰

4. Karau, Holden, Andy Konwinski, Patrick Wendell, and Matei Zaharia. *Learning spark: lightning-fast big data analysis*. " O'Reilly Media, Inc.", 2015.