اولین کنفرانس ملی اصول مهندسی برق و کامپیوتر (PEC-2017)
۶ مرداد ماه ۱۳۹۶، دانشگاه پیام نور استان تهران، تهران، ایران
کد اختصاصی کنفرانس در پایگاه استنادی علوم جهان اسلام (ISC) : ۹۶۱۷۰-۳۱۳۰۲

تحلیل احساس نظرات فیلم‌ها با استفاده از ماشین بردار پیشتیبان دو قلو کمترین مربعات

امیر محمود میر[1*]
[1] دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد تهران شمال
تهران، ایران
mir-am@hotmail.com

سمیه فتاحی[3]
[3]پژوهشگاه علوم و فناوری اطلاعات ایران، ایرانداک
تهران، ایران
fatahi@irandoc.ac.ir

جلال الدین نصیری[2]
[2]پژوهشگاه علوم و فناوری اطلاعات ایران، ایرانداک
تهران، ایران
j.nasiri@irandoc.ac.ir

چکیده — مسئله مورد مطالعه این مقاله تحلیل احساس نظرات فیلم‌ها می‌باشد که تفاوت آن با تحقیقات پیشین استفاده از ماشین بردار پشتیبان دو قلو کمترین مربعات به عنوان دسته بند می‌باشد. از مجموعه داده برگرفته از سایت IMDb برای این تحقیق استفاده شده است که نتایج نشان می‌دهد که روش این مقاله از تحقیقات پیشین بهتر عمل کرده و می تواند نظرات منفی و مثبت راجع به فیلم ها را به خوبی تشخیص دهد. با این حال مسئله تحلیل احساس نظرات فیلم ها چالش های خاص خود را دارد. در پایان پیشنهاد هایی برای تحقیقات آینده مطرح شده است.

کلید واژه — تحلیل احساس، نظرات فیلم ها، ماشین بردار پشتیبان دوقلو کمترین مربعات، پردازش زبان طبیعی

۱. مقدمه

امروزه اطلاعات خیلی زیادی روی اینترنت قرار گرفته است. بخش عمده‌ای از این اطلاعات ساختار مشخصی ندارند که به همین خاطر مدیریت و سازماندهی آنها مشکل شده است. متن کاوی شامل مجموعه ابزار های هوشمندی است که برای سازماندهی اطلاعات بدون ساختار از آن استفاده می‌شود. با کاوش کردن متن ها در اینترنت، ممکن است که به اطلاعات ارزشمندی دست پیدا کنیم [۱].

یکی از سایت های مشهور فیلم، سایت بانک اطلاعات اینترنتی فیلم‌ها[1] می-باشد که یک منبع غنی از اطلاعات راجع به فیلم‌های سینمایی، سریال‌ها و همچنین بازیگران و تهیه‌کنندگان آنها می‌باشد. با وجود فراوان بودن نظرات و نقد ها، یکی از جنبه‌های منفی آن زمان‌بر بودن خواندن این اطلاعات است.

تحقیقات زیادی روی خلاصه سازی متن‌ها صورت گرفته است که بخشی از آنها مربوط به خلاصه سازی نظرات و نقد و بررسی‌ها می‌باشد [۲].

روش‌های ارائه شده در تحقیقات پیشین در حوزه خلاصه سازی نظرات فیلم‌ها، یک نظر را از جنبه احساس بیان شده، به صورت مثبت یا منفی طبقه بندی می‌کنند [۱-۷]. چنانچه نظرات مثبت از منفی بیشتر باشد، کاربر احتمالا تصمیم می‌گیرد که فیلم را تماشا کند.

در این مقاله، مسئله تحلیل احساس نظرات و نقد فیلم‌ها را با رویکرد متفاوت از تحقیقات پیشین، بررسی می‌کنیم. بر اساس اطلاعات ما، در تحقیقات پیشین از SVM[2] استاندارد به عنوان دسته‌بند برای مسئله تحلیل احساس نظرات فیلم‌ها استفاده شده است [۲, ۴]. با این حال، ما در این مقاله از یک مدل توسعه یافته SVM استاندارد به نام ماشین بردار پشتیبان دوقلو کمترین مربعات یاLS-TSVM به عنوان دسته‌بند برای مسئله مقاله استفاده کردیم که نسبت به SVM استاندارد تعمیم پذیری بهتر و حدود ۴ برابر سرعت بیشتری دارد [۸].

۲. تحقیقات پیشین

تحقیقات روی مسئله تحلیل احساس نظرات فیلم‌ها بیش از یک دهه پیش آغاز شده است. در مقاله [۴] یک مجموعه داده برای تحلیل احساس نظرات فیلم‌ها از سایت IMDb ایجاد شد. همچنین از دو دانشجو خواسته شد که ویژگی‌ها یا کلماتی که در نظرات مثبت یا منفی وجود دارد را بیان کنند و با همین ویژگی‌ها اقدام به طبقه بندی احساس نظرات فیلم‌ها کنند که بهترین دقت بدست آمده ۶۹ درصد می‌باشد. در این مقاله از یک تایی‌ها، دوتایی‌ها، برچسب اجزای سخن و جایگاه کلمه در جمله برای استخراج ویژگی استفاده شده است. از بین روش‌های استخراج ویژگی استفاده شده، روش یک تایی ها با SVM استاندارد به دقت ۸۲/۹ رسیده است که از سایر روش‌های استخراج ویژگی بکار گرفته شده، بهتر عمل کرده است.

در مقاله [۱] روش با نظارت یادگیری برای تحلیل احساس فیلم‌ها استفاده شده است به دقت ۸۵/۵۴ درصد دست یافته اند. نتیجه گرفتند که روش‌های با نظارت برای تحلیل احساس نظرات فیلم‌ها مناسب‌تر است. نقطه ضعف این تحقیق عملکرد نامناسب دسته‌بند در تشخیص نظرات منفی است.

[1] Internet Movie Database
[2] Support Vector Machine

در مقاله [۲] از روش SentiWordNet برای تحلیل احساس نظرات فیلم‌ها استفاده شده است. در این روش چهار حالت وجود دارد که یکی از انها اندازه‌گیری احساس یک نظر بر اساس صفات بکار رفته در آن است. به هر صفت در نظر یک امتیاز داده می‌شود که بر اساس مجموع امتیازات، برچسب منفی یا مثبت آن نظر مشخص می‌شود. در این روش برخلاف روش‌های طبقه بندی یادگیری ماشین، نیاز به آموزش نمی‌باشد. با این حال نتایج نشان می دهد که روش‌های طبقه بندی دقت بهتری از روش SentiWordNet دارد.

در تمام تحقیقات پیشین بررسی شده، روش‌های طبقه بندی یادگیری ماشین

$$\begin{bmatrix} w^{(1)} \\ b^{(1)} \end{bmatrix} = -(F^T F + \frac{1}{C1} E^T E)^{-1} F^T e \quad (۴)$$

$$\begin{bmatrix} w^{(2)} \\ b^{(2)} \end{bmatrix} = (E^T E + \frac{1}{C2} F^T F)^{-1} E^T e \quad (۵)$$

برای مسئله تحلیل احساس نظرات فیلم‌ها، توصیه شده است.

۳. چارچوب پیشنهادی تحلیل احساس

۳-۱ پیش پردازش و استخراج ویژگی

مراحل انجام گرفته برای پیش پردازش متن نظرات فیلم‌ها عبارتند از:

۱. ابتدا تمام کلمات توقف از متون حذف شده‌اند.کلمات توقف در متن‌ها فراوان هستند و از این کلمات نمی‌توان به عنوان ویژگی متمایزکننده استفاده کرد.
۲. تمام علامت ها مانند علامت تعجب، نقطه و از این قبیل علائم که در جمله ها زیاد دیده می‌شوند، حذف شده است.

مراحل انجام گرفته برای استخراج ویژگی از متن نظرات فیلم‌ها عبارتند از:

۱. ابتدا تمام یک تایی‌ها در ۲۰۰۰ متن موجود را استخراج کردیم. یک تایی‌ها در این مقاله به عنوان ویژگی شناخته می‌شوند.
۲. بعد از اتمام استخراج یک تایی‌ها، از بین آنها ۲۵۰۰ یک تایی که بیشترین فراوانی را در متن‌ها داشتند، به عنوان ویژگی ها برای تحلیل احساس نظرات فیلم‌ها استفاده شده است.
۳. بعد از اتمام گام دوم، یک ماتریس فراوانی با ابعاد ۲۵۰۰ × ۲۰۰۰ ایجاد می‌شود. ماتریس فراوانی نشان می‌دهد که یک ویژگی در یک متن، چند بار آمده است. با این حال در ماتریس فراوانی تعداد زیادی صفر وجود دارد. برای رفع این مشکل، از شاخص TF-IDF استفاده شده است. معادله ۱ شاخص TF-IDF را نمایش می دهد.

در معادله ۱، متغیر w_{fd} وزن ویژگی f در متن d می‌باشد، متغیر tf_{fd} مقدار فراوانی ویژگی f در متن d است، متغیر D تعداد کل متن ها در مجموعه آموزشی است و همچنین متغیر df_f تعداد متن هایی است که ویژگی f را در خود دارند.

۴. بعد از اتمام گام سوم، یک ماتریس وزن دار برای آموزش دسته بند ایجاد می‌شود که هر سطر آن یک بردار متن می‌باشد.

۳-۲ ماشین بردار پشتیبان دوقلو کمترین مربعات

ماشین بردار پشتیبان دوقلو کمترین مربعات یا LS-TSVM به دنبال دو ابرصفحه غیرموازی است [۸]. بطوریکه روی داده‌های هر کلاس یک ابر صفحه غیر موازی قرار می‌گیرد. برای بدست آوردن این دو ابرصفحه غیر موازی، دو مسئله بهینه سازی با قید تعریف می‌شود که در فرمول ۲ و ۳ نشان داده شده است.

بردار w(i) مختصات ابرصفحه i می‌باشد و b نیز بایاس است. ماتریس A بیانگر داده‌های کلاس مثبت و ماتریس B نیز بیانگر داده‌های کلاس منفی است. بردار e با ابعاد مناسب مقادیر یک را در خود دارد. متغیرy، متغیر لغزش می‌باشد. در فرمول ۲ و ۳ دو پارامتر خطای C1 و C2 وجود دارد که برای رسیدن به کمترین خطا نیاز به تنظیم شدن دارد.

دو مسئله بهینه سازی فوق را می‌توان با دستگاه معادلات خطی حل نمود که راه حل در فرمول ۴ و ۵ آمده است.

ماتریس E برابر است با [A e] و ماتریس F نیز برابر با [B e] می‌باشد. اکنون با فرمول ۴ و ۵ می‌توان مختصات دو ابر صفحه غیر موازی که هر کدام روی یک کلاس قرار می‌گیرند را بدست آورد. برای فهمیدن اینکه داده ارزیابی به کدام کلاس تعلق دارد، فاصله عمودی آن داده را از دو ابرصفحه غیر موازی محاسبه می‌کنیم و سپس داده ارزیابی به کلاسی تعلق دارد که فاصله آن با ابرصفحه غیر موازی آن

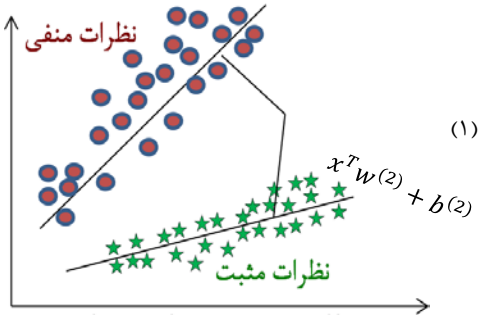
$$Min_{w(1),b(1)} \frac{1}{2}(Aw^{(1)} + eb^{(1)})^T(Aw^{(1)} + eb^{(1)}) + \frac{C1}{2} y^T y \quad (۲)$$
$$\text{s.t.} -(Bw^{(1)} + eb^{(1)}) + y = e$$

$$Min_{w(2),b(2)} \frac{1}{2}(Bw^{(2)} + eb^{(2)})^T(Bw^{(2)} + eb^{(2)}) + \frac{C2}{2} y^T y \quad (۳)$$
$$\text{s.t.} (Aw^{(2)} + eb^{(2)}) + y = e$$

کلاس کمتر باشد.

شکل ۱ برای بهتر توضیح دادن روش LS-TSVM آورده شده است. برای مثال فرض می‌گیریم که نظرات فیلم‌ها دو بعدی هستند و روی کلاس نظرات مثبت و هم منفی، یک خط غیر موازی قرار داده شده است.

$$x^T w^{(1)} + b^{(1)} = 0 \quad (۱)$$

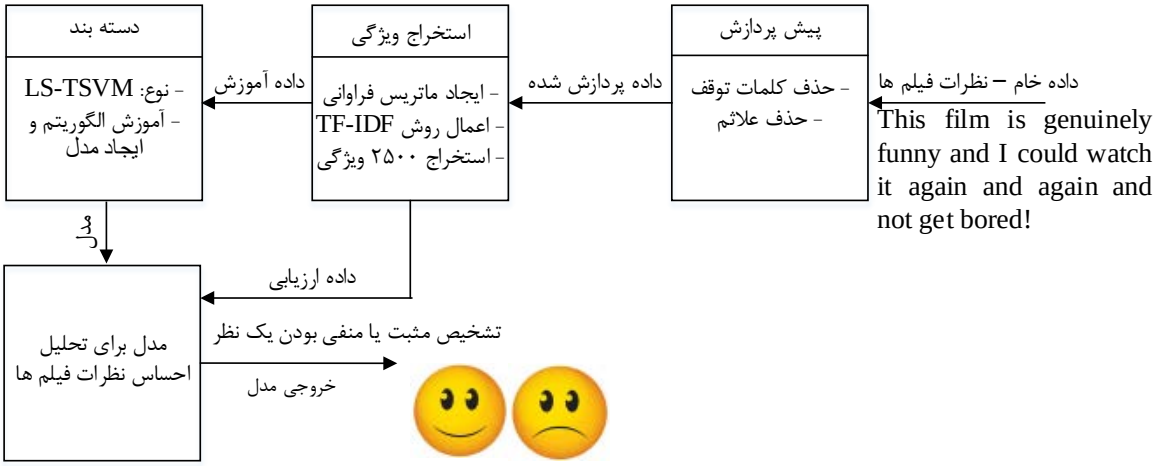


شکل۱: نمایش هندسی روش LS-TSVM

به دلیل محدودیت در تعداد صفحات مقاله نمی‌توانیم جزئیات کامل دسته‌بند LS-TSVM را شرح دهیم. برای اطلاعات بیشتر به مقاله [۸] مراجعه کنید.

۳-۳ روش پیشنهادی این مقاله

در شکل ۲ روش پیشنهادی این مقاله آورده شده است که پیش پردازش، استخراج ویژگی و دسته‌بند LS-TSVM از مولفه‌های روش پیشنهادی هستند. ابتدا داده خام که نظرات فیلم‌ها هستند توسط مولفه پیش پردازش با حذف کلمات توقف و علائم، به یک داده مناسب برای استخراج ویژگی تبدیل می‌شود. سپس مولفه استخراج ویژگی از داده پردازش شده، ۲۵۰۰ ویژگی استخراج می‌کند و داده آموزش را برای دسته بند تولید می‌کند. دسته بند در این مقاله ماشین بردار پشتیبان دوقلو کمترین مربعات می‌باشد که بعد از آموزش، تبدیل به یک مدل برای تحلیل احساس نظرات فیلم‌ها می‌شود. بخشی از داده ها برای ارزیابی مدل ساخته شده، استفاده می‌شود که نتایج این ارزیابی در بخش ۴-۲ آورده شده است. در نهایت خروجی مدل، تشخیص مثبت یا منفی بودن یک نظر درباره یک فیلم است.



شکل۲: مولفه های روش پیشنهادی این مقاله

۴. نتایج آزمایش

۴-۱ مجموعه داده

برای تحلیل احساس نظرات فیلم‌ها، از مجموعه داده منتشر شده توسط IMDb استخراج شده است. این مجموعه داده [9]استفاده شده است که از سایت در هر سه تحقیق اشاره شده در پیشینه تحقیق، آزمایش شده است که نتایج آنها در بخش نتایج آمده است. ویژگی این مجموعه داده یکسان بودن تعداد نمونه های دو کلاس نظرات مثبت و منفی است که از گرایش پیدا کردن دسته بند به یک کلاس جلوگیری می‌کند. تعداد کل نمونه های این مجموعه داده برابر با ۲۰۰۰ نمونه می‌باشد که ۱۰۰۰ نمونه متعلق به کلاس مثبت و ۱۰۰۰ نمونه دیگر متعلق به کلاس منفی می‌باشد. برای دریافت این مجموعه داده می‌توانید به این لینک[3] مراجعه کنید.

۴-۲ نحوه پیاده سازی و اجرای الگوریتم ها

برای پیاده سازی استخراج ویژگی و روش یادگیری ماشین بردار پشتیبان دوقلو کمترین مربعات از زبان برنامه‌نویسی Python[4] نسخه ۳/۶ و همچنین از کتابخانه های Numpy و Scikit-learn استفاده شده است. برای اجرای سایر الگوریتم‌های حاضر در آزمایش مانند درخت تصمیم و بیز ساده از نرم افزار Weka نسخه ۳/۸/۱ استفاده شده است [۱۰].

برای طبقه بندی احساس نظرات فیلم ها از نسخه غیر خطی LS-TSVM با تابع هسته گوسی استفاده شده است که پیدا کردن بهترین پارامترهای آن با جستجوی شبکه‌ای انجام شده است. پارامترهای C1 و C2 بین بازه 2^{-7} تا 2^{12} و پارامتر تابع گوسی نیز بین 2^{-20} تا 2^{4} است. با وجود اینکه روش LS-TSVM از SVM استاندارد حدود ۴ برابر سریعتر است، مدت زمان آموزش دادن دسته‌بند LS-TSVM و پیدا بهترین پارامترها حدود ۶ ساعت زمان برده است. دلیل آن ابعاد بسیار زیاد ماتریس داده‌ها می‌باشد که تعداد ویژگی‌ها برابر با ۲۵۰۰ می باشد.

برای دریافت پیاده‌سازی روش پیشنهادی این مقاله و همچنین مجموعه داده استفاده شده به این لینک[5] مراجعه کنید.

۴-۳ نتایج

در جدول ۱ نتایج تحقیقات پیشین و همچنین این مقاله آورده شده است.

جدول ۱: نتایج آزمایش این مقاله و تحقیقات پیشین

روش یا الگوریتم یادگیری	دقت
ماشین بردار پشتیبان استاندارد – مقاله[۳]	۸۲/۹ درصد

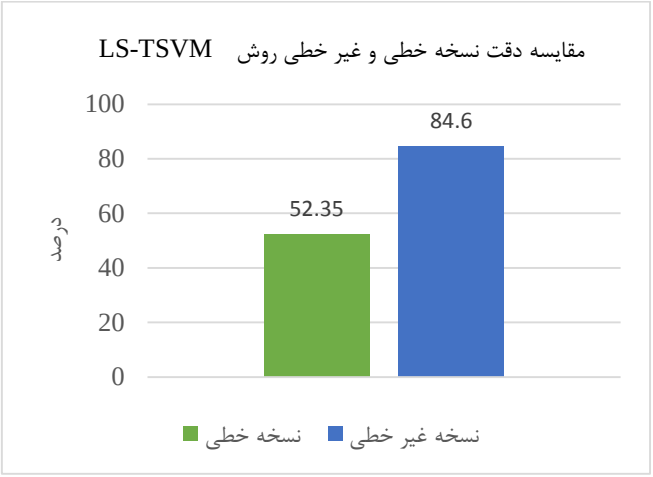
[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[4] http://www.python.org
[5] https://www.dropbox.com/s/xvu3h3ssjy9z08b/SA-Movie-Review-LS-TSVM.rar?dl=0

روش یادگیری با نظارت – مقاله [۱]	۸۵/۵۴ درصد
SentiWordNet – مقاله [۷]	۶۵/۹ درصد
بیز ساده – این مقاله	۷۱/۲ درصد
درخت تصمیم C4.5 – این مقاله	۶۷/۰۵ درصد
LS-TSVM – روش پیشنهادی این مقاله	**۸۴/۶ درصد**

روش بیز ساده و درخت تصمیم با روش استخراج ویژگی این مقاله به ترتیب ۷۱/۲ درصد و ۶۷/۰۵ درصد دقت دارند که نسبت به روش LS-TSVM و ماشین بردار پشتیبان استاندارد، دقت کمتری دارند. بیز ساده و درخت تصمیم از نوع C4.5 با روش استخراج ویژگی بکار گرفته شده در این مقاله، عملکرد خیلی خوبی ندارند. ماشین بردار پشتیبان دوقلو کمترین مربعات به دقت ۸۴/۶ رسیده است که از SVM استاندارد دقت بیشتری دارد که تعمیم پذیری بهتر روش LS-TSVM نسبت به SVM استاندارد را توجیه می‌کند.

در نتایج تحقیقات پیشین، ماشین بردار پشتیبان استاندارد دقت بهتری نسبت به روش SentiWordNet بدست آورده است. بیشترین دقت بدست آمده مربوط به تحقیق [۱] می باشد که با روش یادگیری با نظارت به دقت ۸۵/۵۴ درصد رسیده است. با این حال روش [۱] دارای یک نقطه ضعف مهم می‌باشد که آن هم عملکرد ضعیف این روش روی نمونه‌های کلاس منفی می‌باشد. بطوریکه مقدار بازخوانی[۶] این روش برای کلاس منفی بسیار ضعیف است و نمی‌تواند نمونه های کلاس منفی را به درستی منفی تشخیص دهد.

ماشین بردار پشتیبان دوقلو کمترین مربعات برخلاف روش استفاده شده در مقاله [۱] در تشخیص صحیح نمونه‌های کلاس منفی نیز خوب عمل کرده است. زیرا مجموعه داده استفاده شده یکپارچه می باشد و نمونه های کلاس منفی و مثبت مساوی هستند. به عبارتی در داده های ارزیابی تعداد نمونه‌های کلاس مثبت و منفی تقریبا برابر است و دقت بدست آمده حاکی از تشخیص صحیح نمونه‌های کلاس منفی نیز می‌باشد.



شکل۳: دقت نسخه خطی و غیر خطی روش LS-TSVM

شکل ۳ دقت نسخه خطی و غیر خطی روش LS-TSVM را روی مسئله تحلیل احساس نظرات فیلم‌ها نشان می‌دهد. دقت نسخه غیرخطی LS-TSVM با تابع RBF نسبت به نسخه خطی حدود ۳۲ درصد بیشتر است که بسیار قابل توجه است. همچنین از شکل ۲ می توان برداشت کرد که مسئله تحلیل احساس نظرات فیلم‌ها، یک مسئله غیر خطی جداپذیر می‌باشد.

۵. نتیجه‌گیری و کارهای آینده

در این مقاله از روش استخراج ویژگی مبتنی بر ماتریس فراوانی و اعمال TF-IDF استفاده شده است و دسته‌بند نیز نسخه بهبود یافته ماشین بردار پشتیبان یعنی LS-TSVM می‌باشد. نتایج نشان می‌دهد که روش پیشنهادی این مقاله نسبت به تحقیقات پیشین عملکرد بهتری در مسئله تحلیل احساس نظرات فیلم‌ها دارد و قادر به تشخیص احساس نظرات مثبت و منفی است.

پیشنهادها برای تحقیقات آینده عبارت است از:

۱- در این مقاله، برای تنظیم کردن پارامترهای روش LS-TSVM از جستجوی شبکه‌ای استفاده شده است که زمان آموزش دسته‌بند حدود ۶ ساعت بوده است. با استفاده از الگوریتم‌های تکاملی شاید بتوان زمان آموزش دسته‌بند را به طور قابل توجه‌ای کاهش داد.

۲- در این مقاله، برای تحلیل احساس نظرات فیلم‌ها، از ۲۵۰۰ ویژگی استفاده شده است که ابعاد مسئله را زیاد کرده است. برای تحقیقات آتی می توان از روش‌های کاهش ابعاد مانند PCA برای این مسئله استفاده کرد و نتایج آن را تحلیل و بررسی کرد.

مراجع

[1] Chaovalit, P. and L. Zhou. *Movie review mining: A comparison between supervised and unsupervised classification approaches*. in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. 2005. IEEE.

[2] Zhuang, L. ,F. Jing, and X.-Y. Zhu. *Movie review mining and summarization*. in *Proceedings of the 15th ACM international conference on Information and knowledge management*. 2006. ACM.

[3] Pang, B., L. Lee, and S. Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. 2002. Association for Computational Linguistics.

[4] Kennedy, A. and D. Inkpen, *Sentiment classification of movie reviews using contextual valence shifters.* Computational intelligence, 2006. **22**(2): p. 110-125.

[5] Annett, M. and G. Kondrak. *A comparison of sentiment analysis techniques: Polarizing movie blogs.* in *Conference of the Canadian Society for Computational Studies of Intelligence*. 2008. Springer.

[6] Recall

[6] Ghorbel, H. and D. Jacot, *Sentiment analysis of French movie reviews*, in *Advances in Distributed Agent-Based Retrieval Tools*. 2011, Springer. p. 97-108.
[7] Singh, V., et al. *Sentiment analysis of movie reviews and blog posts*. in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. 2013. IEEE.
[8] Kumar, M.A. and M. Gopal, *Least squares twin support vector machines for pattern classification.* Expert Systems with Applications, 2009. **36**(4): p. 7535-7543.
[9] Pang, B. and L. Lee. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. 2004. Association for Computational Linguistics.
[10] Witten, I.H., et al., *Data Mining: Practical machine learning tools and techniques*. 2016: Morgan Kaufmann.