مدل های زبان روان و سلیس شکلک های اینترنتی برای تحلیل احساسات توییتر

چکیده

در سال های اخیر, تحلیل احساسات توییتر (TSA) به موضوع پژوهشی داغی تبدیل شده است. هدف از پژوهش حاضر, کشف نگرش یا عقیده توییت ها است که نوعاً به عنوان یک مسئله طبقه بندی متن مبتنی بر دانش ماشین تنظیم می شود. برخی روش ها از داده های برچسب گذاری شده دستی برای آموزش مدل های کامل نظارت شده استفاده می کنند, در حالیکه دیگر روش ها از برخی از برچسب های صدادار استفاده می کنند, مانند شکلک های اینترنتی یا هشتگ ها. به طور کلی, ما تنها می توانیم تعداد محدودی از داده های آموزش را برای مدل های کاملاً نظارت شده به دست آوریم, زیرا برچسب گذاری دستی توییت ها بسیار زمانبر و سخت است. در روش های به کار برده شده برای مدل های صدادار, دستیابی به عملکرد رضایت بخش به دلیل نویز (صدای) موجود در برچسب ها سخت است, هرچند دستیابی به میزان زیادی از داده ها برای آموزش آسان می باشد. از اینرو, بهترین راهبرد, استفاده از داده های برچسب گذاری شده دستی و داده های برچسب گذاری شده است. هرچند, نحوه ادغام بی عیب و ایراد این دو نوع متفاوت از داده ها در یک چارچوب آموزش, هنوز یک چالش است. در این مقاله, ما یک مدل جدید را ارائه می دهیم که مدل های زبان روان و سلیس شکلک های اینترنتی (ESLAM) نامیده می شود که این چالش را برطرف خواهد نمود. ایده اصلی این روش, آموزش یک مدل زبانی بر اساس داده های برچسب گذاری شده دستی و سپس استفاده از داده های شکلک اینترنتی صدادار برای روانسازی است. آزمایشات روی مجموعه داده های واقعی نشان می دهد که ESLAM می تواند به طور موثر هر دو نوع داده را برای عملکرد برتر نسبت به هر یک از این روش ها ادغام نماید.

مقدمه

تحلیل احساسات (SA) (Pang and Lee 2007) (که کاوش در عقاید نیز نامیده می شود) عمدتاً در مورد کشف تفکرات دیگران از داده هایی مانند مرور محصول و مقالات خبر است. از یک سو, مصرف کنندگان به دنبال مشورت هایی در مورد یک محصول هستند تا در فرآیند مصرف, تصمیمات آگاهانه بگیرند. از سوی دیگر, فروشندگان بیشتر و بیشتر به عقاید آنلاین در مورد محصولات و خدمات خود توجه می نمایند. از اینرو SA از طرف بسیاری از جوامع پژوهشی مانند آموزش ماشین, داده کاوی و پردازش زبان طبیعی مورد توجه زیادی قرار گرفته است. احساسات در یک متن یا جمله می تواند مثبت, منفی یا طبیعی باشد. ازاینرو, SA در واقع یک مسئله طبقه بندی 3-طرفه است. در عمل, بیشتر روش ها از راهبرد دومرحله ای برای SA استفاده می نمایند (Pang and Lee 2007). در مرحله طبقه بندی ذهنیت, هدف به عنوان ذهنی یا خنثی (عینی) طبقه بندی می شود و در مرحله طبقه بندی قطبیت (تمایل), اهداف ذهنی به صورت مثبت و منفی طبقه بندی می شوند. از اینرو, دو طبقه بندی کننده, برای کل فرآیند SA آموزش می بینند, یکی از آنها طبقه بند ذهنی و دیگر طبقه بند قطبیت نامیده می شود. از زمانی که (Pang, Lee, and Vaithyanathan 2002) SA را به عنوان مسئله طبقه بندی متن بر اساس آموزش ماشین مشخص نمودند, روش های بیشتر و بیشتر آموزش ماشین برای SA پیشنهاد شده است (Pang and Lee 2007).

توییتر یک خدمات وبلاگ نویسی کوچک آنلاین و عمومی است که در سال 2006 راه اندازی شد. کاربران در توییتر, توییت ها را تا حدود 140 کاراکتر می نویسند تا به دیگران بگویند که چه کاری انجام می دهند و چه فکری می کنند. مطابق با گفته برخی منابع, تا سال 2011, بیش از 300 میلیون کاربر در توییتر وجود داشته است و 300 میلیون توییت جدید در هر روز ایجاد می شود. به علت اینکه تقریباً تمام توییت ها عمومی هستند, این داده های فراوان, فرصت های جدیدی را برای انجام پژوهش در مورد داده کاوی و پردازش زبان طبیعی ایجاد نموده است (Liu et al. 2011a; 2011b; 2011c; Jiang et al. 2011).

یک راه برای انجام تحلیل احساسات در توییتر (TSA) بهره برداری مستقیم از روش های SA است (Pang and Lee 2007). هرچند, توییت ها از دیگر اشکال متنی مانند مرور محصول و مقالات خبری متفاوت هستند. اولاً, به

دلیل محدودیت کاراکترها, توییت ها اغلب کوتاه و مبهم هستند. ثانیاً, به دلیل شکل غیرجدی آن, کلمات نادرست املایی, عامیانه و ترکیبی در آن نوشته می شوند. سوماً, میزان زیادی از داده های برچسب گذاری شده صدادار و برچسب گذاری نشده را می توان به آسانی از API توییتر دانلود نمود. بنابراین, بسیاری از روش های SA جدید به طور خاص برای TSA توسعه یافته اند. این روش ها را عمدتاً می توان به دو رده تقسیم بندی نمود: روش های کاملاً نظارت شده و روش های نظارت نشده.

روش های نظارت شده کامل سعی دارند تا طبقه بندی کنندگان را از داده های برچسب گذاری شده دستی آموزش دهند. (Jansen et al. 2009) از مدل Bayes چندجمله ای برای انجام TSA خودکار استفاده نموده اند. (Bermingham and Smeaton 2010) ماشین بردار حمایتی (SVM) و Bayes ساده چندجمله ای (MNB) را برای SA بلاگ و SA میکروبلاگ مقایسه نمودند و دریافتند که SVM در بلاگ هایی با متن های طولانی بهتر از MNB عمل می کند, اما MNB در میکروبلاگ هایی با متون کوتاه بهتر از SVM عمل می کند. یک مشکل در مورد روش های نظارت شده کامل اینست که برچسب گذاری دستی داده ها, زمانبر و سخت است و از اینرو مجموعه داده های آموزش برای بیشتر روش ها اغلب برای تضمین یک کارکرد مناسب بیش از حد کوچک هستند.

کارهای جدیدتر روی روش های نظارت شده از راه دور متمرکز شده اند که با برچسب های صدادار مانند شکلک های اینترنتی و هشتگ ها, طبقه بندها را آموزش می دهد. روش نظارت از راه دور (Go, Bhayani, and Huang 2009) از شکلک های اینترنتی مانند ”(:“ و ”):“ به عنوان برچسب های نویزی برای طبقه بندی تمایل (قطبیت) استفاده می کند. فرض اصلی اینست که یک توییت شامل ”(:“ به احتمال زیاد دارای یک احساس مثبت است و فرض می شود که توییت شامل ”):“ منفی است. آزمایشات نشان می دهند که این شکلک ها حاوی برخی اطلاعات متمایز برای SA می شوند. هشتگ ها (مثلاً #sucks) یا اسمایلی ها در (Davidov, Tsur, and Rappoport 2010) برای شناسایی انواع احساسات استفاده می شوند. (Barbosa and Feng 2010) از داده های صدادار جمع آوری شده از برخی از وب سایت های آشکارسازی احساسات توییتر مانند Twitter

Sentiment3 استفاده نموده اند. (Kouloumpis,Wilson, and Moore 2011) هر دوی هشتگ ها و شکلک ها را بررسی نموده اند و دریافته اند که ترکیب آنها می تواند به عملکرد بهتری نسبت به استفاده از هر یک از آنها منجر شود. مزیت این روش های نظارت شده از راه دور اینست که یادداشت نویسی دستی سخت قابل اجتناب است و میزان زیادی از داده های تعلیم را می توان به راحتی از Twitter API یا وب سایت های موجود ساخت. هرچند, به دلیل نویز (صدا) در برچسب ها, دقت این روش ها رضایت بخش نیست.

با در نظر گرفتن کمبودهای روش های نظارت شده کامل و نظارت شده از راه دور, استدلال ما اینست که بهترین راهبرد, استفاده از داده های برچسب گذاری شده دستی و داده های برچسب گذاری شده صدادار برای تعلیم است. هرچند, نحوه ادغان این دو نوع متفاوت از داده ها در یک چارچوب یادگیری, هنوز یک چالش است. در این مقاله, ما یک مدل جدید, به نام مدل زبان سلیس و روان شکلک های ارنتی (ESLAM) را برای رفع این چالش پیشنهاد می دهیم. مزایای اصلی این روش به شرح زیر می باشند:

- ESLAM از داده های صدادار شکلک ها برای روان نمودن مدل زبان تعلیم یافته از داده های برچسب گذاری شده دستی استفاده می کند. از اینرو این روش, داده های برچسب گذاری شده دستی و داده های برچسب گذاری صدادار در یک چارچوب احتمالاتی را ادغام می کند. میزان زیادی از داده های صدادار شکلک ها که این روش ارائه می دهد, دارای قدرت کار با کلمات سوء تلفظ شده, عامیانه و ترکیبی و کلمات آزمایشی پیش بینی نشده هستند که توسط روش های نظارت شده کامل قابل رفع نیستند.
- در کنار طبقه بندی قطبیت, ESLAM را می توان برای طبقه بندی ذهنیت استفاده نمود که این کار با اغلب روش های نظارت شده موجود از راه دور قابل انجام نیست.
- به جای ریختن میزان زیادی از داده های صدادار در دیسک ها که یک انتخاب معمول توسط روش های نظارت شده موجود از راه دور است, ما یک روش کارآمد و راحت را برای برآورد احتمالات یک کلمه از Twitter API بدون دانلود هر توییت پیشنهاد می کنیم. این کار امیدوارکننده است, زیرا این کار از نظر زمان و ذخیره برای دانلود و پردازش میزان زیادی از توییت ها بسیار گران تمام می شود.

- آزمایشات در مورد مجموعه داده های واقعی نشان داده اند که ESLAM می توان به طور موثر داده های برچسب گذاری شده دستی و داده های برچسب گذاری صدادار در یک چارچوب احتمالاتی را ادغام کند.

کارهای مرتبط

SA (Pang and Lee 2007) دارای سابقه طولانی در پردازش زبان طبیعی است. قبل از (Pang, Lee, and Vaithyanathan 2002)، تقریباً تمام روش ها به طور جزئی مبتنی بر دانش بودند. (Pang, Lee, and Vaithyanathan 2002) نشان داده اند که تکنیک های یادگیری ماشین, مانند Bayes ساده, طبقه بندی های آنتروپی ماکزیمم و SVM می توانند بهتر از روش های پایه مبتنی بر دانش در مرورهای فیلم های سینمایی عمل نمایند. بعد از آن, روش های مبتنی بر یادگیری آموزش به جریان اصلی برای SA تبدیل شده اند. کارهای قبلی روی TSA, از روش های SA سنتی در اشکال متن نرمال مانند مرور فیلم های سینمایی پیروی می نمایند. این روش ها عمدتاً به طور کامل نظارت می شوند (Jansen et al. 2009; Bermingham and Smeaton 2010) که در بخش مقدمه معرفی شده است. بیشتر کارهای اخیر شامل SA وابسته به هدف و مبتنی بر SVM (Jiang et al. 2011)، SA در سطح-کاربرد مبتنی بر شبکه های اجتماعی (Tan et al. 2011)، تحلیل جریان احساسات مبتنی بر قواعد ارتباط (Silva et al. 2011) و SA زمان واقعی (Guerra et al. 2011) می شوند.

اخیراً, روش های نظارت شده راه دور, بیشتر و بیشتر پیشنهاد می شوند. داده های تعلیم (Go, Bhayani, and Huang 2009) شامل توییت ها با شکلک هایی مانند ":)" و ":(" می شوند و آنها از این شکلک های به عنوان برچسب های صدادار استفاده می نمایند. (Davidov, Tsur, and Rappoport 2010) از 50 برچسب توییرت و 15 اسمایلی به عنوان برچسب های صدادار برای شناسایی و طبقه بندی انواع احساسات مختلف توییت ها استفاده نمودند. روش های دیگر با برچسب های صدادار (Barbosa and Feng 2010; Kouloumpis, Wilson, and Moore 2011) نیز پیشنهاد شده اند. تمام این روش ها را نمی توانند طبقه بندی ذهنیت را به

خوبی انجام دهند. علاوه بر این, این روش ها باید تمام داده ها را هدایت کنند و آنها در دیسک ها ذخیره نمایند. زمانی که میلیون ها یا حتی میلیاردها توییت استفاده می شوند, این روش ها نامناسب هستند, زیرا نرخ درخواست برای دادن توییت ها توسط سرور توییتر محدود می شود.

هرچند, بسیاری از روش های TSA پیشنهاد شده اند, تعدادی از آنها می توانند داده های برچسب گذاری شده دستی و داده های برچسب گذاری شده صدادار را در یک چارچوب ادغام نمایند که موجب انگیزش ما برای کار روی ESLAM در این مقاله شده است.

رویکرد ما

در این بخش, ابتدا نحوه انطباق مدل های زبانی (Manning, Raghavan, and Schutze 2009) را برای SA ارائه می دهیم. بنابراین, ما یک روش بسیار کارآمد و موثر را برای آموزش مدل شکلک ها از Twitter API پیشنهاد می دهیم. نهایتاً, ما راهبردی را برای ادغام داده های برچسب گذاری شده دستی و داده های برچسب گذاری شده صدادار در یک چارچوب معرفی خواهیم نمود که روش ESLAM ما می باشد.

مدل های زبانی برای SA

مدل های زبانی (LM) می توانند احتمالاتی یا غیراحتمالاتی باشند. در این مقاله, ما به مدل های زبان احتمالاتی اشاره می کنیم که به طور گسترده در بازیابی اطلاعات و پردازش زبان طبیعی استفاده می شوند (Ponte and Croft 1998; Zhai and Lafferty 2004; Manning, Raghavan, and Schutze 2009).

یک LM, یک احتمال را دنباله ای از کلمات منصوب می کند. در بازیابی اطلاعات, ابتدائاً ما یک LM را برای هر متن تخمین می زنیم, سپس می توانیم یک احتمال را با اندازه گیری این مورد محاسبه نماییم که چقدر احتمال دارد که یک پرس و جو توسط هر LM متنی تولید شود و اسناد با توجه به این احتمالات رتبه بندی شوند.

TSA در واقع یک مسئله طبقه بندی است. برای انطباق LM برای TSA, بر تمام توییت ها از یک رده متمرکز می شویم تا یک متن ساختگی را تشکیل دهیم. از اینرو, برای مسئله طبقه بندی قطبیت, یک نوشته از توییت های آموزشی ممکن ساخته می شود و متن دیگری از توییت های آموزش منفی ساخته می شود. بنابراین ما دو LMها را یاد می گیریم, یکی برای رده مثبت و دیگری برای رده منفی. رویه یادگیری LM برای طبقه بندی ذهنیت مشابه است. در مدت فاز آزمون, ما هر توییت آزمون را به صورت یک پرس و جو در نظر می گیریم و بنابراین می توانیم احتمالات را برای رتبه بندی رده ها استفاده نماییم. یک رده با بالاترین احتمال, به عنوان برچسب برای توییت آزمون انتخاب خواهد شد.

ما c1 و c2 را برای نشان دادن دو مدل زبانی استفاده می نماییم. در طبقه بندی قطبیت, c1, مدل زبان برای توییت های مثبت و c2 برای توییت های منفی است. در طبقه بندی ذهنیت, c1 برای رده ذهنی و c2 برای رده عینی است. به منظور طبقه بندی یک توییت t به صورت c1 یا c2, باید احتمالات توییت را که توسط $P(t|c_1)$ و $P(t|c_2)$ محاسبه می شود, برآورد نماییم. با استفاده از فرض رایج ظهور یک کلمه, خواهیم داشت:

$$P(t|c) = \prod_{i=1}^{n} P(w_i|c),$$

که در آن n تعداد کلمات در توییت t و $P(w_i|c)$ یک توزیع چندجمله ای برآورد شده از LM برای رده c است. این احتمال, فرآیند مولد توییت آزمون را شبیه سازی می کند. اولاً, اولین کلمه (w1) توسط یک توزیع چندجمله ای زیر $P(w_i|c)$ تولید می شود. بعد از آن, دومین کلمه به طور مستقل از کلمه قبلی, با پیروی از همان توزیع تولید می شود. این فرآیند تا زمانی ادامه می یابد که تمام کلمات در این توییت تولید شده باشند.

یک روش رایج مورد استفاده برای برآورد توزیعات, برآورد احتمال ماکزیمم (MLE) است که احتمال را به صورت زیر محاسبه می کند

$$P_a(w_i|c) = \frac{N_{i,c}}{N_c},$$

که در آن $N_{i,c}$, تعداد دفعاتی است که کلمه wi در داده های آموزشی رده c ظاهر می شود و Nc تعداد کل کلمات در داده های آموزشی رده c است.

به طور کلی, لغات توسط مجموعه آموزش تعیین می شوند. برای طبقه بندی توییت های در مجموعه آزمون, روبرو شدن با کلماتی که در مجموعه آموزش ظاهر نمی شوند, بسیار رایج است, به خصوص زمانی که داده ها یا کلمات آموزشی کافی که به خوبی شکل یافته اند, وجود ندارد. در این موارد, روانسازی (Zhai and Lafferty 2004) نقش مهمی در مدل های زبان ایفا می کند, زیرا از انتساب احتمال صفر به کلمات دیده نشده جلوگیری می کند. علاوه براین, روانسازی, مدل را دقیق تر می سازد. روش های روانسازی مشهور شامل روانسازی Dirichlet و روانسازی Jelinek-Mercer (JM) (Zhai and Lafferty 2004) می شوند. هرچند روش روانسازی JM اصلی برای درونیابی خطی مدل MLE با مدل جمع آوری استفاده می شود (Zhai and Lafferty 2004), از روش روانسازی JM برای درونیابی خطی مدل MLE با مدل شکلک ها در این مقاله استفاده می نماییم.

مدل شکلک های اینترنتی

از داده های شکلک های اینترنتی, می توانیم LMها را برای رده های مختلف بسازیم. ما یک روش موثر و کارآمد را برای برآورد LM $P_u(w_i|c)$ شکلک اینترنتی از Twitter Search API پیشنهاد می دهیم. Twitter Search API 4 یک API اختصاص داده شده برای اجرای جستجوها در برابر شاخص زمان-واقعی توییت های اخیر. شاخص آن شامل توییت ها بین 6-9 روز می شود. با توجه به یک پرس و جو که شامل یک یا چند کلمه می شود, API, حدود 1500 توییت مرتبط و زمان ارسال آنها را ارائه می دهد.

طبقه بندی قطبیت برای رسیدن به $P_u(w_i|c_1)$, احتمال wi در رده مثبت, ما یک فرض را می سازیم که تمام توییت های شامل ":)" مثبت هستند. ما یک پرس و جو از ":) wi" را صورت می دهیم و آن را به Search API وارد می کنیم. سپس Search API, توییت های شامل wi و ":)" و زمان ارسال آنها را باز می گرداند. بعد از جمع کردن, به تعداد nwi توییت و گستره زمانی این توییت ها به صورت twi می رسیم. سپس پرس و جوی

دیگری ”(:“ برای صورت می دهیم و به تعداد ns توییت و گستره زمانی ts می رسیم. برخی برآوردها نشان می دهند که یک توییت شامل 15 کلمه به طور میانگین می شوند.

فرض کنید که توییت ها روی توییتر به طور یکنواخت در زمان توزیع شده اند. مشابه با قاعده رسیدن به $P_a(w_i|c)$, می توانیم $P_u(w_i|c_1)$ را با قاعده زیر تخمین بزنیم:

$$P_u(w_i|c_1) = \frac{\frac{nw_i}{tw_i}}{\frac{ns}{ts} \times 15} = \frac{nw_i \times ts}{15 \times tw_i \times ns}.$$

عبارت $\frac{nw_i}{tw_i}$ دقیقاً تعداد دفعاتی است که کلمه wi در رده c در هر زمان ظاهر می شود و عبارت $\frac{ns}{ts} \times 15$ دقیقاً کل تعداد کلمات در رده c در هر واحد زمان است.

در نظر بگیرید که $F_u = \sum_{j=1}^{|V|} P_u(w_j|c)$, ضریب نرمالسازی باشد که در آن $|V|$ اندازه لغات حاوی کلمات دیده شده و دیده نشده است. بنابراین هر $P_u(w_i|c)$ برآورد شده, باید نرمالسازی شوند تا مجموع آنها برابر یک شود:

$$P_u(w_i|c) := P_u(w_i|c)/F_u = \frac{P_u(w_i|c)}{\sum_{j=1}^{|V|} P_u(w_j|c)}$$
$$= \frac{\frac{nw_i \times ts}{15 \times tw_i \times ns}}{\sum_{j=1}^{|V|} \frac{nw_j \times ts}{15 \times tw_i \times ns}} = \frac{\frac{nw_i}{tw_i}}{\sum_{j=1}^{|V|} \frac{nw_j}{tw_i}}.$$

می بینیم که هیچ نیازی به دستیابی به ts و ns نیست, زیرا $P_u(w_i|c)$ را می توان تنها با nwi و twi تعیین نمود.

برای هر LM از رده منفی, فرض می کنیم که توییت های منفی, توییت های شامل ”):“ هستند. رویه تخمین برای $P_u(w_i|c_2)$, مشابه با رویه تخمین برای $P_u(w_i|c_1)$ است. تنها تفاوت اینست که این پرس و جو باید به ”): wi“ تغییر یابد.

طبقه بندی ذهنیت برای طبقه بندی ذهنیت, دو رده, عینی و ذهنی هستند. فرض برای توییت های ذهنی اینست که توییت ها با "(:" یا ")": حامل ذهنیت کاربران هستند. بنابراین ما پرس و جوی (:wi" یا "): را برای رده ذهنی می سازیم.

همانند LM عینی, رسیدن به $P_u(w_i|c_2)$, احتمال wi در رده عینی, چالش برانگیزتر از رده ذهنی است. از نظر ما, هیچ فرض کلی برای توییت های عینی توسط محققان گزارش نشده است. ما راهبردی را امتحان نمودیم که توییت های بدون شکلک را به صورت عینی در نظر می گیرد, اما آزمایشات نشان داد که نتایج رضایت بخش نیستند, که نشان می دهد که این فرض غیرمعقول است. (Kouloumpis,Wilson, and Moore 2011) از برخی از هشتگ ها مانند "#jobs" به عنوان شاخص هایی برای توییت های عینی استفاده نمودند. هرچند, این فرض به اندازه کافی کلی نیست, زیرا تعداد توییت های شامل هشتگ های خاص محدود است و این احساسات توییت ها ممکن است نسبت به موضوعاتی معین مانند jobs (مشاغل) تعصب داشته باشند.

در اینجا ما یک فرض جدید را برای توییت های عینی ارائه می دهیم که فرض می شود که توییت های حاوی یک لینک اینترنتی عینی, باید عینی باشد. بر اساس مشاهده ما, ما دریافتیم که لینک های اینترنتی به سایت های عکس (مثلاً twitpic.com) یا سایت های ویدئویی (مثلاً youtube.com) اغلب ذهنی نیستند و دیگر لینک های اینترنتی مانند لینک های مرتبط با مقالات خبری معمولاً عینی هستند. از اینرو, اگر یک لینک اینترنتی تصویر یا ویدئوها را نشان ندهد, ما آن را یک لینک اینترنتی عینی می نامیم, بر اساس فرض بالا, یک پرس و جو را با عنوان "wifilter : links" 6 برای رسیدن به آمار در مورد رده عینی صورت می دهیم.

ESLAM

بعد از اینکه ما $P_a(w_i|c)$ را از داده های برچسب گذاری شده دستی و $P_u(w_i|c)$ را از داده های شکلک صدادار برآورد نمودیم, می توانیم آنها را در چارچوب احتمالاتی $P_{co}(w_i|c).$ ادغام نماییم. قبل از ترکیب $P_a(w_i|c)$ و $P_u(w_i|c)$, مرحله مهم دیگری وجود دارد: روان نمودن $P_u(w_i|c)$. چون $P_u(w_i|c)$

از داده های شکلک صدادار برآورد می شود, ممکن است تعصب در آن وجود داشته باشد, ما از روانسازی دیریکله برای روانسازی $P_u(w_i|c)$ (Zhai and Lafferty 2004) استفاده نماییم.

با پیروی از اصل روانسازی JM (Zhai and Lafferty 2004), $P_{co}(w_i|c)$ مدل ESLAM ما را می توان به صورت زیر محاسبه نمود:

$$P_{co}(w_i|c) = \alpha P_a(w_i|c) + (1-\alpha)P_u(w_i|c), \qquad (1)$$

که در آن $\alpha \in [0,1]$ پارامتر ترکیب است که سهم هر مولفه را کنترل می کند.

آزمایشات

مجموعه داده ها

Sanders Corpus عمومی دردسترس برای ارزیابی استفاده می شود. که شامل 5513 توییت برچسب گذاری شده دستی می شود. این توییت ها با توجه به یکی از چهار موضوع مختلف جمع آوری شدند (Apple, Google, Microsoft, و Twitter). بعد از حذف توییت های هرز و غیرانگلیسی, 3727 توییت باقیمانده داریم. اطلاعات مفصل از نوشته ها در جدول 1 نشان داده شده است. همانند داده های شکلک صدادار, با نمونه برداری با API, از تمام داده های موجود در توییتر استفاده می کنیم.

Corpus	# Positive	# Negative	# Neutral	# Total
Sanders	570	654	2503	3727

جدول 1: آمار نوشته ها

- ما راهبردهای زیر را برای پیش پردازش داده های اتخاذ می نماییم: نام کاربری. نام های کاربری توییتر که با @ شروع می شود, با "twitterusername" جایگزین می شوند.
- ارقام. تمام ارقام در توییت ها با "twitter digit" جایگزین می شوند.

- لینک ها. تمام لینک های اینترنتی با "twitterurl" جایگزین می شوند.
- کلماتی که بسیار در اینترنت استفاده می شوند. این کلمات مانند the و to حذف می شوند.
- حروف کوچک و کوتاه کردن لغات. تمام کلمات به صورت حروف کوچک نوشته می شوند و کوتاه می شوند.
- توییت دوباره و تکرارها. توییت های مجدد و توییت های تکراری برای اجتناب از وزن اضافی آنها در داده های تعلیم حذف می شوند.

طرح ارزیابی و معیارها

بعد از حذف توییت های مجدد و تکرارها و تنظیم رده ها, به طور تصادفی, 956 توییت را برای طبقه بندی تمایل انتخاب می کنیم که شامل 478 توییت مثبت و 478 توییت منفی می شود. برای طبقه بندی ذهنیت, رده ها را تنظیم می کنیم و به طور تصادفی 1948 توییت را برای ارزیابی انتخاب می کنیم, از جمله 974 توییت ذهنی و 974 توییت عینی.

طرح های ارزیابی برای طبقه بندی ذهنیت و تمایل مشابه هستند. فرض کنید که تعداد کلی توییت های برچسب گذاری شده دستی, شامل داده های آزمون و آموزش, X است. هر بار, به طور تصادفی نمونه را با همان میزان توییت ها (مثلاً Y) برای هر دو رده (مثلاً مثبت و منفی) برای آموزش نمونه برداری کردیم و از X-2Y توییت باقیمانده برای آزمون استفاده نمودیم. این انتخاب تصادفی و آزمون 10 دور به طور مستقل برای هر اندازه مجموعه آموزش منحصر به فرد انجام می شود و عملکرد میانگین گزارش می شود. ما آزمایشات را با اندازه های مختلف از مجموعه آموزش انجام می دهیم, یعنی, Y در مقادیر مختلف مانند 32, 64 و 128 تنظیم می شود.

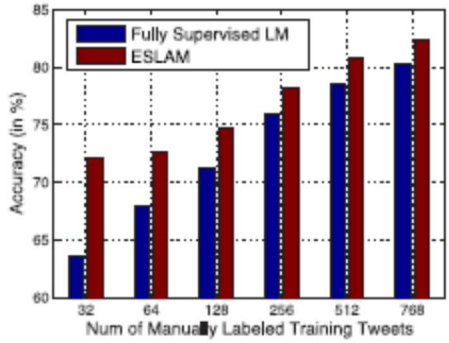
همانند (Go, Bhayani, and Huang 2009) و (Kouloumpis, Wilson, and Moore 2011), دقت و امتیاز-F را به عنوان معیارهای ارزیابی اتخاذ می نماییم. دقت یک معیار از این مورد است که چه درصد از داده های آزمون به درستی پیش بینی می شوند و امتیاز-F توسط ترکیب دقت و یادآوری محاسبه می شود.

اثر شکلک های اینترنتی

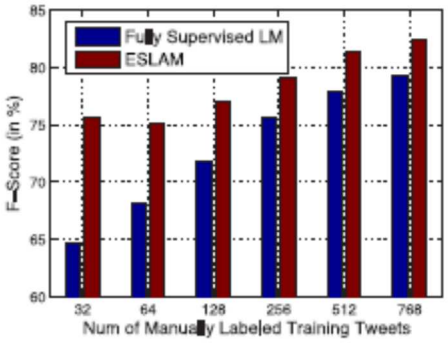
ما روش ESLAM خود را با مدل زبان نظارت شده کامل (LM) مقایسه نمودیم تا تایید نماییم که آیا فصاحت کلام با شکلک ها مفید است یا خیر. لطفاً توجه داشته باشید که LM نظارت شده کامل تنها از داده های برچسب گذاری شده دستی برای تعلیم و آموزش استفاده می کند, در حالیکه ESLAM هر دوی داده های برچسب گذاری شده دستی و داده های شکلک ها را برای آموزش ادغام می کند. شکل 1 و شکل 2 به ترتیب, دقت و امتیاز F را برای دو روش با تعداد متفاوت از داده های تعلیم برچسب گذاری شده دستی نشان می دهد, یعنی $2Y = 32, 64, 128, 256, 512, 768$.

از شکل 1 و شکل 2, می توانیم ببینیم که با افزایش تعداد داده های برچسب گذاری شده دستی, عملکرد دو روش نیز ارتقا می یابد که معقول به نظر می رسد زیرا داده های برچسب گذاری شده شامل اطلاعات متمایز قوی می شوند. تحت تمام تنظیمات ارزیابی, ESLAM بهتر از LM نظارت شده کامل عمل می کند, به خصوص برای تنظیمات با تعداد کمی از داده های برچسب گذاری شده. این نشان می دهد که داده های شکلک صدادار دارای برخی اطلاعات مفید هستند و ESLAM ما می تواند به طور موثر از آن برای دستیابی به عملکرد مناسب بهره برداری نماید.

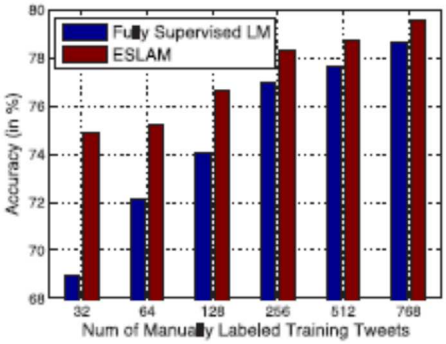
شکل 3 و شکل 4 نشاندهنده دقت و امتیاز-F دو روش روی طبقه بندی ذهنیت با تعداد متفاوت از داده های برچسب گذاری شده دستی است. این نتایج, مشابه با نتایج برای طبقه بندی تمایل هستند که یک بار دیگر, اثربخشی ESLAM را برای استفاده از داده های شکلک صدادار تایید می کند. عملکرد مناسب ESLAM نیز تایید می کند که روش مبتنی بر لینک اینترنتی ما برای یافتن توییت های عینی که یک چالش بزرگ برای بیشتر روش های نظارت شده از راه دور موجود می باشد, مفید است.



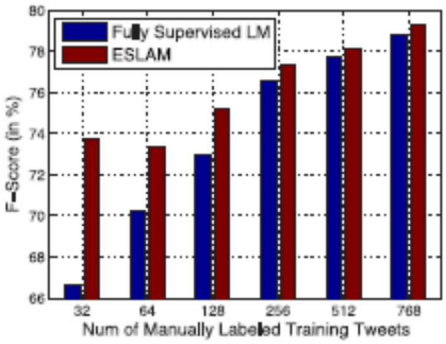
شکل 1: اثر شکلک های اینترنتی روی دقت طبقه بندی قطبیت



شکل 2: اثر شکلک های ارنتی روی امتیاز-F طبقه بندی قطبیت



شکل 3: اثر شکلک های اینترنتی روی دقت طبقه بندی ذهنیت



شکل 4: اثر شکلک های اینترنتی روی امتیاز-F طبقه بندی عینیت

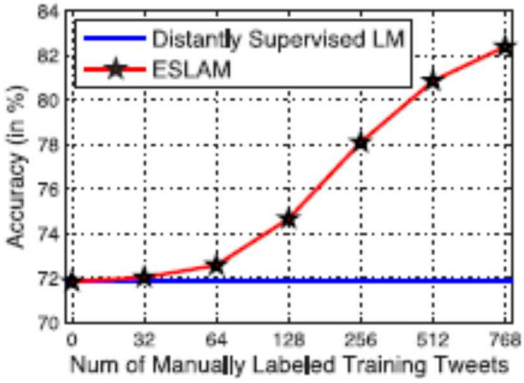
اثر داده های برچسب گذاری شده دستی

ما روش ESLAM خود را با LM نظارت شده مقایسه نمودیم تا تایید نماییم که آیا داده های برچسب گذاری شده دستی می توانند اطلاعات مفیدتر را برای طبقه بندی فراهم نمایند یا خیر. لطفاً توجه داشته باشید که LM نظارت شده تنها از داده های صدادار شکلک ها برای آموزش استفاده می کند, در حالیکه ESLAM از داده های برچسب گذاری شده دستی و داده های شکلک برای آموزش استفاده می کند.

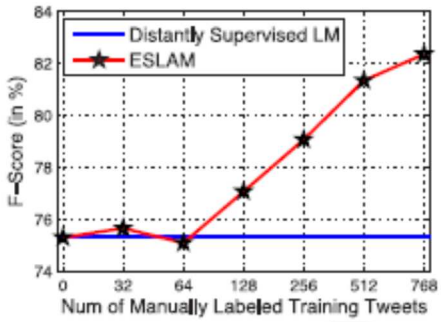
شکل 5 و شکل 6, دقت و امتیاز-F دو روش را روی طبقه بندی قطبیت با تعداد متفاوت از داده های برچسب گذاری شده دستی نشان می دهد.

خط آبی متناظر با عملکرد LM نظارت شده است که همچنین متناظر با مورد داده های برچسب گذاری شده دستی صفر است. خط قرمز, نتایج ESLAM است. می توانیم ببینیم که ESAM, به عملکردی بهتر از LM نظارت شده دست می یابد. با افزایش داده های برچسب گذاری شده دستی, تفاوت عملکرد بین آنها بیشتر و بیشتر می شود. بنابراین این ادعای ما تایید می شود که استفاده از داده های برچسب های صدادار برای آموزش به تنهایی کافی نیست.

شکل 7 و شکل 8, دقت و امتیاز-F دو روش را روی طبقه بندی ذهنیت با تعداد متفاوت از داده های برچسب گذاری شده دستی آموزش نشان می دهد. این نتایج با طبقه بندی قطبیت مشابه هستند.



شکل 5: اثر داده های برچسب گذاری شده دستی بر دقت طبقه بندی قطبیت (تمایل)



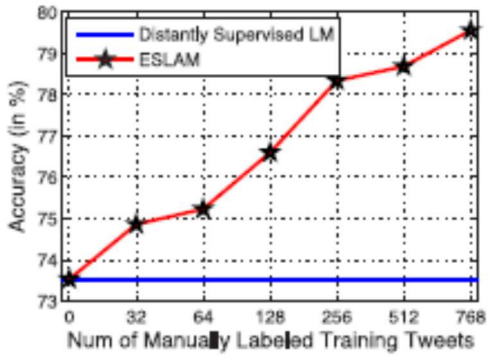
شکل 6: اثر داده های برچسب گذاری شده دستی بر امتیاز-F قطبیت (تمایل)

حساسیت به پارامترها

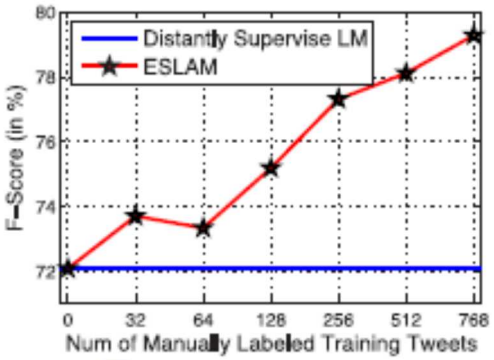
پارامتر α در (1), نقشی حیاتی در کنترل تسهیم بین اطلاعات برچسب گذاری شده دستی و اطلاعات برچسب گذاری شده صدادار ایفا می کند. برای نشان دادن اثر این پارامتر به صورت مفصل, مقادیر مختلف را برای طبقه بندی قطبیت امتحان می کنیم. شکل 9 و شکل 10, دقت ESLAM را به ترتیب با 128 و 512 توییت نشان می دهند.

مورد $\alpha = 0$ بدان معنیست که تنها داده های شکلک صدادار استفاده می شوند و $\alpha = 1$ مورد کاملاً نظارت شده است. نتایج در اشکال به وضوح نشان می دهند که بهترین راهبرد, ادغام داده های برچسب گذاری شده دستی و داده های صدادار در تعلیم است. ما همچنین خاطر نشان می نماییم که با 512 داده تعلیم برچسب گذاری شده,

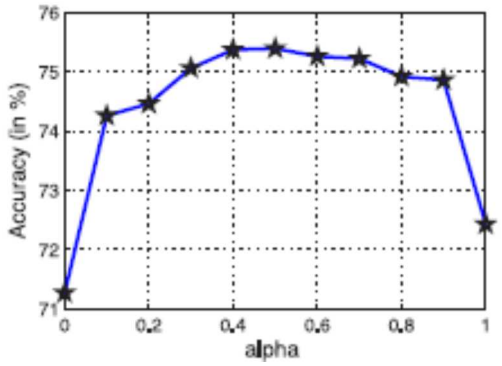
ESLAM به بهترین عملکرد با α بزرگتر از مورد 128 داده برچسب گذاری شده دست می یابد که به طور مشخص معقول می باشد. علاوه بر این, ما در می یابیم که ESLAM به تغییرات کوچک در مقدار پارامتر α حساس نیست, زیرا گستره α برای دستیابی به عملکرد بهتر بزرگ است.



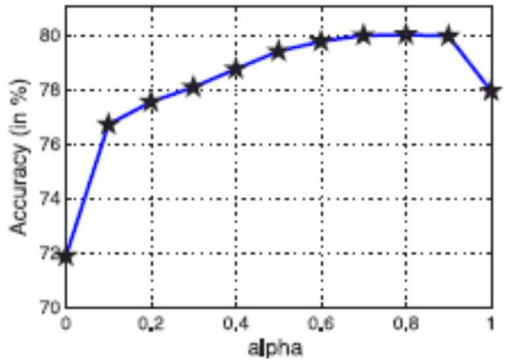
شکل 7: اثر داده های برچسب گذاری شده دستی بر دقت طبقه بندی ذهنیت



شکل 8: اثر داده های برچسب گذاری شده دستی بر امتیاز-F طبقه بندی ذهنیت



شکل 9: اثر پارامتر روانسازی α با 128 توییت آموزش برچسب گذاری شده



شکل 10: اثر پارامتر روانسازی α با 512 توییت آموزش برچسب گذاری شده

نتیجه گیری

روش های موجود از داده های برچسب گذاری شده دستی و داده های برچسب گذاری شده صدادار برای تحلیل احساسات در توییتر استفاده می کنند, اما تعداد کمی از آنها از هر دو برای آموزش استفاده می نمایند. در این مقاله, ما یک مدل جدید به نام مدل زبان سلیس شکلک های اینترنتی (ESLAM) برای ادغام این دو نوع داده در یک چارچوب احتمالاتی را پیشنهاد نمودیم. آزمایشات روی مجموعه داده های واقعی نشان می دهند که روش ESLAM ما می تواند به طور موثر هر دو نوع داده را برای عملکرد بهتر از هر یک از این روش ها ادغام نماید.

روش ESLAM ما, به اندازه کافی برای ادغام انواع دیگر از برچسب های صدادار برای آموزش مدل که در کارهای آنیده دنبال خواهد شد, کلی و عمومی می باشد.

References

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING*, 36–44.

Bermingham, A., and Smeaton, A. F. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *CIKM*, 1833–1836.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*, 241–249.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *Technical report*.

Guerra, P. H. C.; Veloso, A.; Jr., W. M.; and Almeida, V. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *KDD*, 150–158.

Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *JASIST* 60(11):2169–2188.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *ACL*, 151–160.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*, 538–541.

Liu, X.; Li, K.; Zhou, M.; and Xiong, Z. 2011a. Collective semantic role labeling for tweets with clustering. In *IJCAI*, 1832–1837.

Liu, X.; Li, K.; Zhou, M.; and Xiong, Z. 2011b. Enhancing semantic role labeling for tweets using self-training. In *AAAI*.

Liu, X.; Zhang, S.; Wei, F.; and Zhou, M. 2011c. Recognizing named entities in tweets. In *ACL*, 359–367.

Manning, C. D.; Raghavan, P.; and Schutze, H. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

Pang, B., and Lee, L. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, 79–86.

Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *SIGIR*, 275–281.

Silva, I. S.; Gomide, J.; Veloso, A.; Jr., W. M.; and Ferreira, R. 2011. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *SIGIR*, 475–484.

Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M.; and Li, P. 2011. User-level sentiment analysis incorporating social networks. In *KDD*, 1397–1405.

Zhai, C., and Lafferty, J. D. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2):179–214.