شیوه نامه برچسب گذاری

محدودیت های مجموعه داده های فارسی

مجموعه داده مورد استفاده در روش‌های مشابهت سنجی معنایی به صورت جفت جمله برچسب دار می‌باشند که با عددی در بازه‌ی (0و1) یا با یکی از اعضای مجموعه {0,1,2,3,4,5} میزان مشابهت دو جمله یا سند بیان می‌شود. با توجه به دانش ما، تاکنون تمامی پروژه‌های شباهت سنجی معنایی فارسی بر روی مجموعه داده های سرقت ادبی و با هدف کشف موارد تطابق متن صورت گرفته است. این نوع مجموعه داده ها شامل متون رسمی و آکادمیک و اسناد با طول بلند می‌باشند. تاکنون پژوهشی معنامحور بر روی متون کوتاه عامیانه شبکه‌های اجتماعی مانند داده‌های توئیتر در زبان فارسی انجام نشده است. در نتیجه مجموعه داده ای برای این امر در حال حاضر وجود ندارد. در این پژوهش برای تست و ارزیابی مدل پیشنهادی، نیاز به مجموعه داده ای برچسب گذاری شده داریم، که با توجه به عدم وجود چنین مجموعه داده‌ای، برآن شدیم تا مجموعه داده ای از داده‌های توئیتر فارسی به منظور تستِ مدل شباهت سنج معنایی ایجاد و در آزمایشات خود استفاده کنیم.

همانطور که پیش‌تر ذکر شد، مجموعه داده‌ی وظایف شباهت سنجی معنایی به صورت جفت-سند هم معنی با برچسبی به منظور نمایش امتیاز مشابهت دو سند است. یافتن توئیت‌های هم‌معنی نیازمند شناسایی پارافریزها می‌باشد. بسیاری از کارهای قبلی در متون انگلیسی در مورد شناسایی پارافریز بر روی یک مجموعه داده خاص، مجموعه پارافریز مایکروسافت [52]، که از مقالات خبری مشتق شده است، توسعه و ارزیابی شده است. اما این داده‌ها با داده‌های توییتر بسیار متفاوت است. در توئیتر اخبار بسیار کوتاه درج می‌شوند و به دلیل محدودیت در تعداد کاراکتر، بیان مطالب توسط واژگان خاص انجام می‌شود. با شناسایی این واژگان مشترک بین توئیت های ارسالی در یک موضوع و یک زمان خاص، می‌توان توئیت های پارافریز را شناسایی کرد.

در کارهای گذشته در زبان‌های غیر فارسی مانند انگلیسی، عربی و ترکی [53] [54] [55]، روال کار به این صورت است که بعد از شناسایی توئیت‌های پارافریز، توسط چند حاشیه‌نوس خبره جفت توئیت ها از لحاظ معنایی

مقایسه می‌شوند و امتیازی به عنوان برچسب دریافت می‌کنند. درنهایت با توجه به رای حاشیه‌نویس‌ها برچسب میانگین شناسایی شده و به عنوان امتیاز هر جفت توئیت در مجموعه داده درج گردیده است.

جمع‌آوری و برچسب‌زنی مجموعه داده تست

در این بخش، مراحل گردآوری و برچسب‌گذاری مجموعه داده جمع آوری شده با هدف مشابهت سنجی معنایی داده های شبکه‌های اجتماعی، به ترتیب شرح داده می‌شود. لازم به ذکر است، مجموعه داده مورد استفاده این پژوهش مبتنی بر داده های جمع آوری شده فارسی توییتر در بستر سامانه ذکاوت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) بوده و همچنین جهت پردازش زبان فارسی نیز از ابزارهای توسعه داده شده در این سامانه استفاده خواهد شد. به صورت کل پس از جمع‌آوری موضوعی داده ها در تاریخ های مشخص، با کمک روش بخش 5-3-1 پارافریز های موجود شناسایی و برای عملیات برچسب زنی ذخیره و پالایش می شوند. دیتاست جمع آوری شده در مجموع 1123 جفت-توئیت را شامل می شود که توسط 4 حاشیه نویس مجرب برچسب زنی شده اند. برچسب نهایی هر نمونه از دیتاست، با میانگین گیری از برچسب این 4 نفر بدست می آید.

شناسایی جفت داده‌ی کاندید

در ابتدا نیاز است به این موضوع پرداخت که دیتاست های مربوط به وظیفه ی مشابهت سنجی معنایی، به صورت جفت (pair) می باشند. جفت متن به این صورت که دو متن از لحاظ معنایی به یکدیگر مشابهت دارند (دو متن پارافریز هم هستند) و این مشابهت با مقدار عددی به عنوان درجه تشابه تحت عنوان برچسب در دیتاست ذخیره شده اند. پس هر نمونه در دیتاست از یک جفت متن با برچسبی که نشان دهنده میزان مشابهت دو متن است، نمایش داده می شود.

برای ایجاد جفت های کاندید، ابتدا تمام داده ها پیش پردازش شدند و مواردی که به مقایسه‌ی معنایی متون کمکی نمی‌کردند (برای مثال stopwords، هشتگ های غیرمرتبط و ...) را حذف کردیم. در مرحله بعد، تمام جفت متون به عنوان جفت‌های کاندید در نظر گرفته شدند، سپس با توجه به معیارهای زیر فیلتر شدند. یک جفت کاندید حذف می شود اگر تنها یکی از شروط زیر برقرار باشد:

1. در یک جریان (trend) از شبکه ی اجتماعی مورد نظر نباشند.
2. فاصله زمانی درج دو متن بیشتر از یک روز باشد.
3. دو متن در کمتر از 5 لغت همپوشانی داشته باشند.

امتیازدهی شباهت

روش حاشیه نویسی ما از روشی پیروی می کند که در اکثر دیتاست های زبان انگلیسی برای مشابهت سنجی معنایی استفاده می شود. به جای اینکه جفت متون را صرفا به‌عنوان پارافریز یا غیرپارافریز برچسب‌گذاری کنیم، یک امتیاز تشابه معنایی دقیق‌تر به آن اختصاص داده می‌شود. ما دستورالعمل‌های ارائه‌شده برای وظیفه‌ی شباهت سنجی معنایی را دنبال کردیم و حاشیه‌نویس‌ها جفت متون را مستقیما، طبق معیارهای نشان‌داده‌شده در جدول زیر نمره‌گذاری کردند.

نمرات تشابه معنایی برای جفت پارافریزهای کاندید

درجه مشابهت معنایی	راهنما
5 - یکسان	کاملاً معادل هستند، زیرا آنها به یک معنا هستند.
4- نزدیک	به مقدار زیادی معادل هستند، اما برخی از جزئیات بی اهمیت متفاوت هستند.
3- مرتبط	تقریبا معادل هستند، اما برخی از اطلاعات مهم متفاوت است/از دست رفته است.

2- در یک زمینه	معادل نیستند، اما برخی جزئیات را به اشتراک می‌گذارند.
1- تا حدودی مرتبط	معادل نیستند، اما در یک موضوع هستند.
0- غیر مرتبط	در موضوعات مختلف هستند.

چهار گویشور بومی فارسی زبان، به عنوان حاشیه نویس استخدام شدند. ما مجموعه ای از دستورالعمل های حاشیه نویسی را برای توضیح روند امتیازدهی آماده کردیم. این دستورالعمل، نمرات شباهت را به همراه یک جفت مثال برای هر نمره و توضیح کوتاه معرفی کرده است. نمونه‌ها نیز از میان نمونه‌های متعدد انتخاب شدند.

به منظور استفاده بهتر و آزمایش‌های دقیق تر در آینده، نمرات اختصاص داده شده نیز به برچسب های باینری تبدیل شدند. ابتدا، نمرات هر حاشیه نویس با در نظر گرفتن جفت متون که به صورت یکسان (برچسب 5)، نزدیک (برچسب 4) و مرتبط (برچسب 3) به عنوان مثبت (یعنی پارافریز) و آنهایی که در یک زمینه (برچسب 2)، مرتبط (برچسب 1) و غیرمرتبط (برچسب 0) علامتگذاری شده بودند، برچسب منفی هستند (یعنی غیر پارافریز). تعداد تصمیمات مثبت و منفی برای هر نمونه ممکن است به صورت یک جفت خلاصه شود. به عنوان مثال، (1,3) نشان می دهد که تنها یک حاشیه نویس این جفت را به عنوان یک پارافریز برچسب گذاری کرده است، در حالی که سه نفر باقی مانده آن را به عنوان غیرپارافریز برچسب گذاری کرده اند. در جدول 5، معیارهای قضاوت باینری را نشان می دهیم که بر اساس تعداد پاسخ های حاشیه نویسان است. توجه داشته باشید که در مواردی که تعداد تصمیمات مثبت و منفی برابر باشد، یک جفت را **"قابل بحث"** (debatable) در نظر می‌گیریم. این رویکرد مشابه روش برچسب گذاری دیتاست TPC است که بر اساس توافق بین حاشیه نویسان نیز می باشد.

برای برچسب نهایی شباهت معنایی، ما میانگین نمرات را ارائه می دهیم. این محدوده بین 1.75 و 3.00 برای جفت های "قابل بحث" است، در حالی که جفت های پارافریز مثبت بالاتر از 3.00 و جفت های منفی کمتر از 1.75 امتیاز دارند.

معیارهای قضاوت باینری بر اساس تعداد پاسخ های حاشیه نویس ها

تعداد پاسخ ها	قضاوت
(4,0); (3,1)	مثبت (1)
(0,4); (1,3)	منفی (0)
(2,2)	قابل بحث

برای مثال جدول شماره 6، سه جفت نمونه از داده ها را نشان می دهد. هر جفت با نمرات 4 حاشیه نویس مختلف و میانگین نمرات شباهت نشان داده شده است. جفت "قابل بحث" توسط چهار حاشیه نویس (4,2,3,0) نمره گذاری شده است و میانگین شباهت 2.25 است.

جفت های نمونه فرضی از دیتاست

جفت متن	امتیازات	پارافریز
تاجر سرشناس در مورد مافیای خودرو گفت: "اگر برخی چیزها را توضیح دهم، آنها نمی توانند در ایران بمانند" ن.ح. تاجر ایرانی در گفتگو با انتخاب 24 گفت: "مافیای خودرو بگذراند سکوت کنم، وگرنه جایی در این کشور ندارند"	(4,2,3,0) $Average = 2.25$	قابل بحث
مشخص شد که 13 خودرو از 15 خودرو که در قرعه کشی به یک فرد رسید، در نمایندگی های منطقه هشت سپرده شده است.	(3,4,5,5) $Average = 4.25$	مثبت

13 شانس از 15 خودرو که به یک شخص رسید، در منطقه هشت است که از حساب دلالان منطقه واریز شده است.		
استان فارس با 52.9 درصد بیشترین سهم فروش وام مسکن در کل فروش مسکن را داشت. سهم اولین فروش از کل فروش خانه 45.4 درصد بوده است.	$(1,3,0,0)$ $Average = 1$	منفی

چند نمونه از برچسب دهی داده ها

برای درک بهتر و آشنایی بیشتر با نحوه‌ی لیبل دهی توسط حاشیه نویس ها، این بخش به ذکر چند مثال از دیتاست و توضیح مختصری در مورد آن می‌پردازد.

نمونه 1 :

توئیت یک	توئیت دو	امتیاز مشابهت
نیکلاس #مادرو رئیس جمهوری #ونزوئلا وارد #تهران شد.	نیکلاس مادرو، رئیس جمهور ونزوئلا وارد تهران شد	5

همانطور که در نمونه شماره یک دیده می‌شود، دو توئیت کاملا از لحاظ معنایی و محتوایی یکسان می‌باشند و اندک تفاوتی از لحاظ لغوی در آنها مشاهده می‌شود. پس در نتیجه دو توئیت پارافریز یکدیگر بوده و با توجه به جدول شماره 1، با امتیاز مشابهت 5 می‌توان آن‌ها را برچسب‌دهی کرد.

نمونه2:

توئیت یک	توئیت دو	امتیاز مشابهت

دیدار دوستانه ایران - عمان به خاطر مخالفت برانکو منتفی شد با وجود توافق فدراسیون‌های فوتبال ایران و عمان و نامه‌نگاری‌هایی که برای برگزاری دیدار دوستانه انجام شده بود، اما این دیدار به دلیل مخالفت برانکو ایوانکوویچ منتفی شد	#برانکو هم به درخواست #تیم_ملی ایران جواب منفی داد با وجود توافق فدراسیون‌های فوتبال #ایران و عمان و نامه‌نگاری‌هایی که برای برگزاری دیدار دوستانه انجام شده بود، اما این دیدار به دلیل مخالفت برانکو ایوانکوویچ منتفی شد.	5

از آنجایی که در نمونه شماره دو نسبت به شماره یک تفاوت های لغوی بیشتری دیده می شود، اما با بررسی معنایی دو توئیت موجود در این نمونه، مشاهده می‌شود که محتوای دو متن کاملا یکسان است و تمامی معانی موجود در توئیت یک، در توئیت شماره دو نیز موجود است، از این رو امتیاز مشابهت 5 برای این جفت در نظر گرفته می‌شود.

نمونه 3:

توئیت یک	توئیت دو	امتیاز مشابهت
در ایران خبرگزاری صدا و سیما می‌گوید نایب رئیس فدراسیون فوتبال به اتهام کلاهبرداری بازداشت شده. گفته شده این دستگیری با یک پرونده کلاهبرداری با استفاده از نوعی رمز ارز مرتبط است. گفت‌وگو با فرید اشرفیان، خبرنگار ورزشی:	نایب‌رئیس فدراسیون فوتبال بازداشت شد نایب‌رئیس فدراسیون فوتبال به‌دلیل ارتباط با پرونده #کلاهبرداری رمزارز جعلی کینگ‌مانی بازداشت شد	4

در نمونه شماره 3، دو توئیت از لحاظ معنا و محتوا به یک موضوع اشاره دارند، اما در توئیت یک به مواردی کم اهمیت دیگری پرداخته شده که به بیان اصل موضوع ارتباط چندانی ندارد. به همین دلیل امتیاز مشابهت 4 برای این جفت لحاظ می‌شود.

نمونه 4:

توئیت یک	توئیت دو	امتیاز مشابهت
مقام جهاد کشاورزی: ظرف یک هفته قیمت برنج را به نرخ مصوب برمی‌گردانیم	با افزایش قیمت #برنج ایرانی به کیلویی ۱۰۰ هزار تومان، جهاد کشاورزی از اجرای طرح تشدید مبارزه با احتکار برنج در ایران خبر داد و اعلام کرد: بر اساس این طرح، قیمت برنج ظرف یک هفته به نرخ مصوب باز خواهد گشت. قیمت‌های دستوری در چهار دهه گذشته جزیی از سیاست اقتصادی در #ایران بوده است	3

در نمونه 4، موضوع مورد بحث در هر دو توئیت به یک چیز اشاره می‌کند، اما برخی از اطلاعات مهم در توئیت یک وجود ندارد و یا اینکه متفاوت از توئیت دو است. به همین دلیل امتیاز مشابهت 3 در نظر گرفته می‌شود.

نمونه 5:

توئیت یک	توئیت دو	امتیاز مشابهت

سردار آزمون بهترین شد #سردار_آزمون با گرفتن نمره بسیار خوب 8.4 از وب‌سایت آماری هواسکورد بهترین بازیکن لورکوزن در بازی با گروتر فورث شد	عکس روز - اولین گل #سردار_آزمون در #بوندسلیگا. آزمون که از ابتدا در ترکیب تیم لورکوزن حضور داشت، موفق شد در دقیقه ۱۸ بازی مقابل گروترفورث گل دوم تیمش را به ثمر برساند.	2

در نمونه شماره 5، از لحاظ معنایی و محتوایی دو توئیت به دو موضوع متفاوت اشاره می‌کنند، اما در برخی جزئیات مشترک هستند، در نتیجه مرتبط نبوده اما می‌توانیم با استفاده از جدول شماره 1، امتیاز مشابهت 2 را به آن‌ها اختصاص دهیم.

نمونه 6:

توئیت یک	توئیت دو	امتیاز مشابهت
پیکر بی‌جانی که خانواده #عبدالباقی دارن بالای سرش اون نمایش مضحک و بی‌معنی رو بازی میکنن، پیکر عزیز یک خانواده است که احتمالاً در نزدیکی آوار #متروپل بی‌قرار و دل‌آشوب منتظر خبری	تست DNA و دم خروس #عبدالباقی ۱- تستDNA نیاز به یک تا سه هفته زمان برای کپی‌برداری و مقایسه دارد و در بهترین آزمایشگاههای دنیا حداقل۳ روز زمان نیاز دارد و اعلام نتیجه ۲۸ ساعت بعد از پیداشون جسد تامل برانگیز است ۲-تستDNA مقایسه دو	1

هستند؛ حقیقتاً قساوت و پستی این جماعت انتها نداره	DNA با هم است. روی DNA آدم‌ها اسمشون نوشته نشده	

دو توئیت نمونه شماره 6، از لحاظ معنایی پارافریز هم نبوده و ارتباطی با یکدیگر ندارند، اما در یک موضوع مشترک قرار دارند و به همین علت امتیاز مشابهت 1 برای آن‌ها لحاظ شده است.

نمونه 7:

توئیت یک	توئیت دو	امتیاز مشابهت
چرا و چرا و چرا #عبدالباقی در ساختمانی که مشکل دارد دفتر میسازد! شاید توهم داشته مقاوم است. یعنی اینقدر به خودش دروغ گفته که باورش شده.	«برادارن لیلا» برنده جایزه «فیپرشی» #جشنواره_کن شد هیات داوران فدراسیون بین‌المللی منتقدان (فیپرشی) در جشنواره فیلم کن، جایزه بهترین فیلم بخش رقابتی اصلی را به فیلم «#برادران_لیلا» ساخته #سعید_روستایی اعطا کرد. روستایی جایزه‌اش را به مردم داغ‌دیده #آبادان تقدیم کرد_انتخاب	0

نمونه شماره 7، حاوی دو توئیت از دو موضوع متفاوت و کاملا جدا از هم است که باید امتیاز مشابهت 0 را برای آن‌ها در نظر گرفت.

	tweet 1	tweet 2	نفر اول	نفر دوم	نفر سوم	نفر چهارم	average
1	ترجمه: یکی از هدیه هایی که ترامپ می تواند به کشور و جهان بدهد این است که در ۱۵ روز آخر ریاست جمهوری خود جنگی با ایران آغاز نکند. دشوار هست بگوییم این کار عمدا از طرف دولت صورت میگیرد یا اینکه به طور ساده نتیجه بی کفایتی آنهاست. در هر صورت ، آنها در حال انجام یک بازی خطرناک هستند. https://t.co/sDJySiyMSW	@Moghadam۹۶۳۱ ((مشخص نیست شاید این صوت از طرف خودشون باشه)) ▼هشدار هیلاری کلینتون: ترامپ، با ایران جنگ نکن!این یک بازی خطرناک است ◆هیلاری کلینتون در پیام توییتری توشته: هدیه‌ای که ترامپ می‌تواند به این کشور و جهان بدهد این است که در ۱۵ روز باقی‌مانده از دولتش جنگی با ایران آغاز نکند. https://t.co/Ghdvdwv۷Yeb	4	3	5	4	4
2	#حسن_یزدانی استوری گذاشته و قبل از عذر خواهی بابت باخت دیشب به خانواده #مهسا‌امینی تسلیت گفته. قهرمان و پهلوان قلب ما خودتی حسن آقا یزدانی	به نظر من حسن یزدانی یک قهرمان عالی و تکرار نشدنی هست ولی آقا حسن پهلوان نیست پهلوان باید پشت مردمش باشه لطفا برداشت بد نکنید 🙏 #بختی #حسن_یزدانی https://t.co/۲THPltX۴gq	2	4	4	3	3.25
3	«خداروشکر با قتل پسرم به آرامش رسیدم/دیگر هیچ ناراحتی در زندگی ندارم/نمیدانستم پسرم در خانه ما رابطه ی جنسی برقرار میکند» اگه زیاد دلشو ندارید این رشتو رو نخونید. بخشی از اعترافات پدر #بابک_خرمدین:	پدر #بابک_خرمدین : پسرم دختران زیادی را به خانه می آورد و در اتاق با آنها تنها بود. ما فکر میکردیم با آنها رابطه جنسی برقرار میکند. از رفتارهایش خسته شدیم و اورا کشتیم. «متهم در هنگام خروج از اتاق بازپرس دستهایش را رو به آسمان کرد و خدا را بابت	4	4	4	4	4

سه نمونه موجود از دیتاست گردآوری شده این پژوهش

شکل زیر سه نمونه از دیتاست گردآوری شده را نشان میدهد.

برای دریافت اطلاعات بیشتر از پروژه با ایمیل zekavat@itrc.ac.ir تماس بگیرید.