عنوان پایاننامه

کشف زیر رخداد در جریان داده خبری

چکیده پایان نامه

حجم تولید اطلاعات در سطح جهان روندی تصاعدی و شگفت‌انگیز را طی می‌نماید. استفاده از اطلاعات درصورتی‌که به تولید و ارائه دانش و دانائی منتهی گردد، می‌تواند دستاوردهای زیادی را به دنبال داشته باشد. یکی از ویژگی‌های مهم عصر اطلاعات، میزان تولید، ذخیره‌سازی و نشر اطلاعات در جهان است و یکی از انواع داده‌ها، متون خبری تولیدشده توسط رسانه‌ها است. حجم عظیم داده‌های خبری تولیدشده توسط رسانه‌های مختلف، اهمیت روش‌هایی برای دسته‌بندی موضوعی، تجمیع خبرهای مشابه ازلحاظ محتوایی و ارائه رخدادهای جدید به کاربر به‌صورت آنی، دقیق، پالایش‌شده و بر اساس نیاز کاربر را آشکار می‌کند.

اکثر رخدادها دچار تغییر و تحول می‌شوند و بنابراین، درون هر رخداد ممکن است تعداد زیادی زیر رخداد شکل گیرد که وضعیت جاری رخداد را توصیف می‌کند. رخداد عبارت است از رویدادی که در زمان و مکان مشخص اتفاق می‌افتد، کشف رخداد جدید نیز به معنی کشف اولین سند خبری درباره یک رخداد خاص در میان جریانی از اسناد خبری پیوسته است که از لحاظ زمانی مرتب شده‌اند. زیر رخداد را می‌توان یک واحد مجزا به لحاظ معنایی از یک رخداد کامل دانست که بخشی از یک رخداد کلی را تشکیل داده و در یک بازه زمانی کوتاه موردبحث قرارگرفته و سپس، زیر رخداد جدید شکل‌گرفته و یا رخداد کلی خاتمه می‌یابد. هدف این پژوهش، کشف زیر رخداد با در اختیار داشتن مجموعه‌ای از اخبار مرتبط به یک رخداد معین است. روش پایه برای کشف زیر رخداد جدید، محاسبه شباهت متنی خبر جاری با خبرهای قبلی و تصمیم‌گیری درباره زیر رخداد جدید و یا تکراری بودن بر اساس یک مقدار آستانه است. چالشی که در روش پایه وجود دارد، وجود خبری است که دارای محتوای جدید است اما به دلایلی، با مجموعه اخبار قبلی دارای شباهت بالایی بوده که این موضوع منجر به خطای تشخیص خبر جدید خواهد شد. چالش دیگر، تنوع واژگانی در متون خبری است، بدین معنی که خبری با محتوای تکراری، دارای شباهت کم با مجموعه اخبار قبلی بوده و در نتیجه، خطای تشخیص خبر تکراری رخ می‌دهد. در این پژوهش تلاش شده است که توانایی روش شباهت معنایی بین اسناد، در حل چالش تنوع واژگانی مورد مطالعه قرار گیرد. علاوه بر این، با تقسیم‌بندی متن خبر و استفاده از مدل زبانی موقعیتی، چالش شباهت بالای متنی نیز مورد مطالعه و آزمایش قرارگرفته است. برای حل چالش های مطرح شده، تاثیر ویژگی زمان انتشار خبر، شباهت خبرگزاری و داده های تولیدشده توسط کاربران در شبکه‌های اجتماعی و نیز، نظرات کاربران در بخش نظرات خبر، مورد بررسی قرار گرفته است.

به منظور انجام آزمایش‌ها، مجموعه اخبار از سه خبرگزاری تابناک، خبر آنلاین و فردانیوز جمع‌آوری شده، اخبار پیرامون چند رخداد خاص تجمیع و سپس به صورت دستی برچسب زده شده است. آزمایش‌هایی که روی این مجموعه داده ای انجام شده است نشان می‌دهد که استفاده از مدل زبانی موقعیتی و فاصله معنایی در مقایسه با روش پایه به مراتب موفق‌تر عمل کرده است. همچنین، تاثیر استفاده از داده های تولیدشده توسط کاربران، ویژگی زمان انتشار خبر و شباهت خبرگزاری برای حل دو چالش مطرح شده نیز مورد بحث قرارگرفته و نتایج آن ارائه شده است.

کلمات کلیدی
کشف زیر رخداد،اخبار فارسی، شباهت معنایی، مدل زبانی موقعیتی، مدل ترکیبی بازخورد، شباهت کسینوسی نرم، فاصله کلمات موور

کلمات کلیدی انگلیسی
Sub-event detection, semantic similarity, positional language model, mixture model feedback, soft cosine similarity, word mover distance