با ظهور وب ۲٫۰ و ایجاد تالارهای گفتگو کاربران اینترنت این قابلیت را پیدا کردند تا مستنداتی در فضای اینترنت ایجاد کنند و با یکدیگر به بحث و تبادل نظر در مورد موضوعات مختلف بپردازند. از آنجایی که حجم زیادی از این مستندات ایجاد شده توسط کاربران حاوی نظرات آنها در مورد یک موضوع، رویداد یا یک محصول است تحلیل این نظرات می‌تواند سودمند باشد. به عنوان مثال مطالعات نشان می‌دهد که اشخاص برای خرید یک محصول به نظرات کاربران به نسبت رسانه‌های دیگر چون تبلیغات و کاتالوگ‌ها توجه بیشتری دارند، و از آنجایی که خواندن این حجم از اطلاعات برای کاربران مقدور نیست، وجود یک سیستم که به صورت خودکار نظرات کاربران را تحلیل و رده‌بندی کند مورد نیاز است. با ظهور شبکه‌های اجتماعی نظیر توییتر که کاربران پست‌هایی متنی با حداکثر ۱۴۰ کاراکتر می‌توانند منتشر کنند و میل بالای کاربران به انتشار نظرات خود در این شبکه‌ها در سالهای اخیر محققین تلاش‌هایی برای نظرکاوی مستندات این شبکه‌ها انجام داده‌اند.

در ابتدا بیشتر تلاش‌ها در حوزه توییتر متمرکز بر استفاده از روش‌های سنتی بود. اما حالت محاوره‌ای توییتر تعیین قطبیت اسناد را با چالش‌هایی همچون کلمات مخفف، غلط‌های املایی و کنایه‌آمیز بودن اسناد مواجه کرده است. یکی دیگر از چالش‌ها تنوع در موضوعات مورد بحث در توییتر است که باعث می‌شود کلمات در موضوعات متفاوت قطبیت متفاوتی داشته باشند.

در این پژوهش سعی شده چالش‌های تعیین قطبیت اسناد توییتری را مورد مطالعه قرار دهیم و روش مبتنی بر مدل زبانی را که پیش از این معرفی شده به گونه‌ای تغییر دهیم که این چالش‌ها مرتفع گردد. ابتدا با پیش‌پردازش‌های صورت گرفته تا جای ممکن تنک بودن فضای کلمات را کاهش می‌دهیم. سپس با استفاده از ایجاد مدل موضوعی سعی می‌کنیم مدل‌های زبانی ایجاد شده را با توجه به موضوع سند بهبود دهیم و در خاتمه برای غلبه بر چالش فضای تنک کلمات با استفاده از نمایش برداری کلمات قطبیت کلمات جدید را با استفاده از تشابه برداری با کلماتی که قبلا در داده آموزش دیده شده است تخمین می‌زنیم. با توجه به نتایج به دست آمده از آزمایشات استفاده از مدل موضوعی برای تشخیص موضوع سند و همینطور استفاده از تشابه برداری کلمات برای تعیین قطبیت کلمات به تازگی دیده شده در داده آزمون می‌توانند بر این چالش‌ها غلبه کنند و نتایج بهتری به‌دست دهند.