

Resumen

Los políticos y los ciudadanos están utilizando las redes sociales de forma intensiva para expresar sus opiniones sobre los debates públicos. Si bien por una parte las interacciones en la web permiten a los usuarios conocer diferentes puntos de vista, por otra parte no resuelven conflictos, por el contrario, a menudo contribuyen a polarizar aún más el debate. Las plataformas de redes sociales como Twitter proporcionan a los investigadores una gran cantidad de contenido generado por los usuarios. Con lo que explorar la opinión pública e investigar cómo las personas se comunican entre sí es ahora posible como nunca antes. El interés por identificar las opiniones expresadas en los textos generados en las redes sociales ha aumentado significativamente en los últimos años. Se han propuesto métodos de procesamiento de lenguaje natural basados en algoritmos de aprendizaje automático o enfoques de aprendizaje profundo para detectar la opinión de los usuarios sobre un tema específico discutido en las redes sociales. Además, varios trabajos han sugerido que en las redes sociales existe la segregación ideológica, aunque estas herramientas pueden potencialmente exponer a los usuarios a una gama más amplia de puntos de vista diferentes. En esta tesis abordamos el problema de la detección de las opiniones en las redes sociales, centrándonos en los debates políticos polarizados en Twitter. La detección de opiniones (stance detection) consiste en determinar automáticamente si el autor de una publicación está a favor o en contra de un objetivo de interés, o si no se puede inferir la opinión. Nos ocupamos de temas políticos como las elecciones políticas y los referendos y, como resultado, los objetivos son tanto personas como referendos. También exploramos las comunicaciones que tienen lugar en estos debates polarizados, arrojando luz sobre las dinámicas de comunicación entre personas que tienen opiniones en acuerdo o en conflicto, enfocándonos en particular en la observación del cambio de opiniones (opinion shifting). Proponemos modelos de aprendizaje automático para la stance detection como si fuera un problema de clasificación binaria. Exploramos características basadas en el contenido del texto del tweet, además usamos características basadas en información contextual que no emerge directamente del texto. Utilizando el corpus de benchmark propuesto para la tarea compartida sobre la stance detection realizado para SemEval 2016, exploramos la contribución que el estudio de las relaciones entre el objetivo de interés y las otras entidades involucradas en el debate proporciona a la stance detection. En particular, consideramos a los dos candidatos principales que competían para las elecciones primarias del Partido Demócrata y Republicano antes de las elecciones presidenciales de los Estados Unidos de 2016. Los resultados superan los obtenidos por los equipos que participaron en la tarea. Nuestro modelo hace uso del conocimiento de las relaciones entre el objetivo de interés y las entidades mencionadas para inferir la opinión, incluso cuando el objetivo no es mencionado directamente. Al participar en la tarea 'Stance and Gender Detection in Tweets on Catalan Independence' organizado para IberEval 2017, hemos propuesto otras características textuales y contextuales para la stance detection en tweets en español y en catalán. Nuestro sistema (iTACOS) consiguió la primera posición entre los diez equipos participantes para ambos idiomas en la subtarea de stance detection. Explorando la stance detection desde una perspectiva multilingüe, hemos creado un corpus de tweets en francés y uno en italiano. Hemos decidido seleccionar temas que sean muy similares a los dos corpus de benchmark publicados en SemEval 2016 e IberEval 2017 para que los nuevos conjuntos de datos sean más comparables a los ya existentes. El corpus francés (E-FRA) consta de tweets sobre la segunda vuelta de las elecciones presidenciales francesas del 2017. Para la creación del corpus (R-ITA) hemos recogido tweets italianos sobre el referéndum constitucional italiano. La extensión multilingüe de nuestro modelo (multiTACOS) muestra que la stance detection está influenciada más por los diferentes estilos utilizados por los usuarios para comunicar la opinión sobre objetivos de diferentes tipos (personas o referendos) en lugar del idioma utilizado. Con el objetivo de recuperar información contextual sobre la red social de los usuarios de Twitter (generalmente las tareas compartidas solo consisten en el contenido del tweet, dejando de lado la información sobre el usuario), hemos creado otros dos conjuntos de datos, uno en inglés y uno en italiano, respectivamente, sobre el Brexit (TW-BREXIT) y sobre el referéndum constitucional italiano (ConRef-STANCE-ita). En ambos casos de estudio, mostramos que los

usuarios tienden a agruparse en grupos con ideas similares. Por este motivo, el modelo que explota el conocimiento de la comunidad social a la que el autor del tweet pertenece, supera los resultados obtenidos utilizando solo las funciones basadas en el contenido de la publicación. Además, la evidencia muestra que los usuarios utilizan diferentes tipos de comunicación según el nivel de acuerdo con la opinión del interlocutor, por ejemplo, las relaciones de amistad, los retweets y las citas (quote) son más comunes entre los usuarios relacionados, mientras que las respuestas (replies) se utilizan a menudo para interactuar con usuarios que tienen diferentes posiciones. Al abordar la stance detection desde una perspectiva diacrónica, también observamos tanto el cambio de opinión como la mitigación del debate hacia posiciones neutrales después del resultado de la votación. Además, hemos observado que tener contacto con una variedad más amplia de opiniones puede influir en la propensión a cambiar de opinión. Finalmente, mostramos que las características basadas en una representación gráfica de un dominio de interés no se limitan a la stance detection, sino que se puede aplicar a diferentes escenarios. Al proponer otra tarea de clasificación que realiza la identificación del talento en el deporte, especialmente en el estudio de caso del tenis de mesa, mostramos que las métricas de redes basadas en la centralidad son una señal fuerte para el talento y pueden usarse para entrenar un modelo de algoritmo de aprendizaje automático para enfrentar esta tarea.