**Unidad 1**

**Tarea 1**

**David Aarón Ramírez Olmeda**

**A case study of Spanish text transformations for twitter sentiment analysis**

**Resumen**

El objetivo de esta investigación es identificar en un amplio conjunto de combinaciones qué transformaciones de texto (lematización, stemming, eliminación de entidades, entre otras), tokenizadores (por ejemplo, n-gramas de palabras), y esquemas de ponderación de tokens tienen el mayor impacto en la precisión de un clasificador (Support Vector Machine) entrenado en dos conjuntos de datos españoles. Los resultados muestran que esta novedosa combinación de palabras y caracteres produce un clasificador que supera a la combinación tradicional basada en palabras en un 11,17% y un 5,62%, respectivamente.

y un 5,62% en los conjuntos de datos INEGI y TASS'15, respectivamente.

Además, introducimos un nuevo enfoque basado en la combinación de n-gramas basados en palabras y q-gramas basados en caracteres. La metodología utilizada consiste en analizar exhaustivamente todas las combinaciones de transformaciones de texto y sus respectivos parámetros para averiguar qué características comunes tienen los clasificadores con mejores resultados.

**Introducción**

Dado que el análisis de texto es una herramienta útil para saber lo que piensa la gente sobre cualquier cosa; así, representa un importante apoyo para los procesos de toma de decisiones (para cualquier nivel de gobierno, marketing, etc.) Estos hechos han tenido como consecuencia que se vean los medios sociales como una mina de oro de opiniones humanas y, en consecuencia, el interés por realizar investigaciones y actividades empresariales en torno a la minería de opiniones y el análisis de sentimientos.

En los últimos años, la producción de documentos textuales en los medios sociales ha aumentado exponencialmente. Por ejemplo, hasta abril de 2016, Twitter cuenta con 320 millones de usuarios activos, y Facebook con 1590 millones de usuarios.

El análisis automático del sentimiento de los textos es una de las tareas más importantes de la minería de textos, en la que el objetivo es determinar si un documento concreto tiene una opinión positiva, negativa o neutra. En las redes sociales, la gente comparte comentarios temas dispares, como eventos, personas y organizaciones, entre otros.

entre otros.

**Desarrollo**

De forma contraintuitiva, descubrimos que la complejidad de las técnicas utilizadas en la fase de preprocesamiento no está correlacionada con el rendimiento final del clasificador; por ejemplo, un clasificador que utilice la lematización, que es una de las técnicas de preprocesamiento con mayor complejidad, podría no ser uno de los sistemas con mayor rendimiento.

El segundo enfoque utiliza un conjunto de reglas de grano fino para aprovechar la información léxica; se diseñaron aproximadamente 50 reglas teniendo en cuenta los marcadores de negación. Lematización (lem) El proceso de lematización es una tarea compleja del Procesamiento del Lenguaje Natural que determina el lema de un grupo de formas de palabras, es decir, la forma de diccionario de una palabra. Además, las palabras coloquiales y las abreviaturas se transforman utilizando una expresión regular basada en un diccionario de ese tipo de palabras.

Además, también proporcionamos el análisis experimental de la combinación de tokenizadores, que mejora el rendimiento global sin alejarse demasiado de nuestra estructura de clasificador de texto. En las siguientes subsecciones, describimos las técnicas de transformación de texto utilizadas, que tienen su contrapartida en muchos idiomas, pero cuya correcta implementación depende en gran medida del idioma de destino, en nuestro caso el español. Comparación del rendimiento de n-palabras y q-gramas.

En palabras grandes, se capturan los sufijos y prefijos; cuando q es pequeña, también se capturan los afijos y la raíz de la palabra. En primer lugar, hay que transformar el texto en una representación adecuada, lo que se conoce como modelización de textos. El análisis automático del sentimiento de los textos es una de las tareas más importantes de la minería de textos, cuyo objetivo es determinar si un documento concreto tiene una opinión positiva, negativa o neutra3.

Además, proporcionamos una lista de reglas sencillas para encontrar un clasificador de sentimiento basado en nuestros hallazgos; no obstante, la mejor configuración depende del conjunto de datos, las clases y muchas otras propiedades dependientes de la tarea. El texto del lado izquierdo de la flecha se sustituye por el texto del lado derecho.

Por último, queremos dejar claro que utilizamos SVM como clasificador debido a su popularidad en la comunidad, este trabajo se centra principalmente en el tratamiento del texto independientemente, por lo que la selección adecuada y el ajuste del clasificador se deja como trabajo futuro.

**Conlusión**

Dimos una lista de reglas sencillas para encontrar un clasificador de sentimientos basado en nuestros resultados. La mejor configuración depende del conjunto de datos, las clases y muchas otras propiedades dependientes de la tarea. Por último, dedejamos claro que utilizamos SVM como clasificador debido a su popularidad en la comunidad.

En este trabajo, nuestro enfoque consiste en realizar una evaluación exhaustiva del espacio de parámetros y luego ampliar la búsqueda utilizando una combinación de tokenizadores.

**The Data-Information-Knowledge-Wisdom Hierarchy and its Antithesis**

**Resumen**

Los errores pueden producirse incluso cuando los datos o la información son exactos, a través de datos o información precisos, mediante una mala interpretación o un razonamiento erróneo. No proceden de la falta de datos, sino de la distorsión o comunicación incorrecta de información o el conocimiento, incluidos la información y el conocimiento correcto ede la fuente.

**Introducción**

En Deadly Decisions, Christopher Burns (2008) examina los antecedentes del hundimiento del Titanic, la explosión del transbordador espacial Challenger y otras catástrofes, descubriendo que en todos los casos se ignoraron las advertencias, las salvaguardias del sistema diseñadas para alertar a los operadores sobre una crisis bloquearon o disimularon datos reales, los sistemas de comunicación se rompieron, y la culpa e incluso el error fueron negados y encubiertos por todas las partes. Esto produce los siguientes resultados:

Datos Ausencia o falta de datos; datos que faltan; Información Desinformación; desinformación; error; Conocimiento Ignorancia; Sabiduría Locura; estupidez

**Desarrollo**

Obsérvese que no existe un término único para lo contrario de datos.

La forma jerárquica de concebir las relaciones entre datos la información, el conocimiento y la sabiduría engendrada por la jerarquía DIKW de Ackoff puede servir un marco interesante para considerar lo opuesto al conocimiento como un espectro, y esto nos permite circunscribir un dominio que podemos llamar no-conocimiento, que plantea

opuestos de cada uno de los términos de la jerarquía DIKW. La perspectiva de Ackoff

de la jerarquía, de modo que para él "comprender" significa la capacidad de evaluar y corregir errores, mientras que "sabiduría" significa la capacidad de evaluar cualquier acto y evaluarlas en relación con el ideal de control total (omnicompetencia).

A pesar de las objeciones de que el modelo pasa por alto los distintos tipos de información, conocimiento y sabiduría, este modelo jerárquico y piramidal capta de personal y organización, sino también para la organización de los recursos bibliotecarios. Pero la noción de interconectar estos fenómenos como procesos que conducen de los datos brutos a la sabiduría a través de la filtración, la reducción y el refinamiento, creó una imagen poderosa de gran valor en la enseñanza y la reflexión de las ciencias de la información y las bibliotecas de gran valor para la enseñanza de la biblioteconomía y la documentación y para la reflexión.

La jerarquía DIKW puede arrojar luz sobre los problemas de la organización del conocimiento. Un error es una equivocación cometida inadvertidamente o por ignorancia, mientras que la desinformación es una información errónea o un falso falsa de la inteligencia recibida, y la desinformación es la difusión deliberada de información falsa con la intención de influir en las opiniones e incluso en las políticas de quienes la reciben. Una investigación hacia una teoría de este tipo se pregunta, en efecto, qué reside, si es que existe algo, fuera del universo del conocimiento documental, cómo lo describiríamos, cómo accederíamos a él, cómo organizarlo en las bibliotecas y cómo identificarlo para distinguirlo del conocimiento válido.

Dado el DIKW, el conocimiento está conectado en un sistema de fases interconectadas e y verbalmente identificadas que representan diferentes niveles de complejidad, síntesis y refinamiento, es necesario un concepto que reúna todas las contrapartidas negativas de estos procesos y fenómenos. Del mismo modo, un nuevo libro titulado The Virtues of Ignorance (Las virtudes de la ignorancia), editado por Vitek y Jackson (2008), reclama una

de nuestra cultura basada en el conocimiento e inspirada en la Ilustración, que nos permita admitir nuestra ignorancia casi total como especie.

**Conclusión**

Aunque la mayoría de los psicólogos y otros escritores se han ocupado durante mucho tiempo de distinguir la estupidez de la discapacidad mental, el significado real de la estupidez no está claramente definido en relación con la locura. en relación con la locura. La jerarquía Datos-Información-Conocimientos-Sabiduría como pirámide, all crear su jerarquía, Ackoff se centró únicamente en modos específicos de datos, información, conocimiento y sabiduría, y descuidó importantes distinciones observadas por científicos de la información como Buckland (1991), Machlup (1980) y Soergel (1985).

El tema ha sido abordado recientemente por Damian Thompson en Counterknowledge (2008), que abarca la ciencia de la creación, la medicina alternativa, los esquemas de enriquecimiento rápido, las historias en los libros de texto afrocéntricos y los sitios web que promulgan teorías conspirativas sobre los atentados terroristas del 11 de septiembre.