Informe del Proyecto Final de Sistemas de Recuperación de Información

Ernesto Bárcena Trujillo, José Antonio Concepción Alvarez, y Alejandro García González

Universidad de La Habana, La Habana, Cuba
3rnst020421@gmail.com
concepcijoseantonio@gmail.com
juanpereiraojosrojos@gmail.com

Resumen El uso de internet se ha incrementado en los últimos años, con este aumento las personas han utilizado distintas plataformas para ofrecer su opinión sobre productos, servicios, personas, etc. La sobrecarga de estas opiniones a menudo dificulta que los consumidores identifiquen reseñas útiles a través de las funciones tradicionales. El objetivo de este trabajo es implementar un algoritmo para obtener un mejor sistema de clasificación.

Palabras Claves: Clasificación de reseñas · Procesamiento de Lenguaje Natural · Grandes Modelos de Lenguaje · Python

1. Introducción a la problemática

El contenido de las reseñas generado por un usuario es útil para que los compradores tomen decisiones más informadas en sitios web de compras en línea. Sin embargo, debido al gran volumen de reseñas, especialmente de artículos populares, es muy difícil para cualquiera encontrar fácilmente información relevante. A menudo, el sitio web sólo proporciona algunos métodos de clasificación simples, pero no muy útiles, como clasificación por tiempo o calificación por votos de utilidad. Se pueden considerar varios factores para este propósito, incluido el historial de compras reciente del usuario y el de navegación o las propiedades de los artículos. Cualquier solución para este problema implica un análisis profundo de los datos ofrecidos por los usuarios (en lenguaje natural), así como, el uso de información contextual de un caso más específico del problema, para poder extraer ideas relevantes tanto para los clientes como para los compradores. Las implementaciones se pueden encontrar en este proyecto en GitHub.

1.1. Procesamineto del lenguaje hasta hoy

Los primeros trabajos en procesamiento del lenguaje natural (PLN) se centraron en tareas básicas como la traducción automática y la extracción de información, aunque la comprensión profunda del significado del texto era un objetivo lejano. Durante décadas, los investigadores desarrollaron sistemas basados

en reglas, definiendo manualmente extensos conjuntos de reglas gramaticales y semánticas para analizar el texto. A finales del siglo XX, se produjo un cambio hacia métodos estadísticos, utilizando grandes corpus de texto para entrenar modelos probabilísticos que aprendieran automáticamente las regularidades del lenguaje. En los últimos años, el aprendizaje profundo, especialmente a través de redes neuronales recurrentes (RNN) y "Transformers", ha revolucionado el campo, logrando resultados sorprendentes en tareas como la traducción automática, la generación de texto y la respuesta a preguntas.

2. Soluciones propuestas

2.1. Algoritmo determinista

Soluciones del estado del arte

- aqui mencionar un poco de lo que se ha hecho con referencias a algunos papaers

Primeras aproximaciones sin aprendizaje automático

Esbozadas estas ideas, planteamos la siguiente solución para ofrecer más conocimientos acerca de la información de los comentarios. Podemos ver cada comentario como la descripción de un objeto o servicio, donde se detalla la experiencia del usuario o se proporciona simplemente una clasificación sobre una de las características del ítem. A estas características las llamaremos atributos o "features". La solución propuesta estará basada en la extracción de atributos de un ítem con su respectiva caracterización. Luego, estos serán evaluados según la positividad de los mismos, con el objetivo de confeccionar un ranking fidedigno a la positividad de cada ítem. Finalmente, según los intereses de un usuario, se podría sugerir un mejor comentario más alineado con sus preferencias. Además se tendrá en cuenta el uso de otras métricas para medir la calidad de los textos así como información de valor social en el proceso de retroalimentación de las comunidades.

Para la extracción de atributos existen diversas variantes; entre ellas, podemos utilizar la biblioteca NLTK, la cual posee diversas herramientas para el análisis de texto. Una forma de extraer los atributos es analizar la ocurrencia de sustantivos y adjetivos en un texto. Para obtener mejores resultados, es necesario separar cada comentario por oraciones, lo cual, asumiendo que el texto está bien estructurado (esta métrica se abordará más adelante con mayor detalle), nos permitirá realizar una mejor extracción. El algoritmo para la extracción consiste en, dado un sustantivo forma oracional de sustantivo, extraer todos los adjetivos que lo modifican, teniendo en cuenta que estos deben estar en una misma oración. Para la extracción de adjetivos se utiliza un criterio simple de cercanía la sustantivo.

Veamos el siguiente ejemplo de un comentario sobre un hotel.

"El hotel era maravilloso. La comida estaba decente. Las personas que trabajaban ahí eran muy buenas. La playa sensacional."

En este caso se extraen los pares:

- hotel maravilloso
- comida decente
- personas buenas
- playa sensacional

De esta forma podemos obtener atributos del objeto en cuestión así como una descripción de ellos. Gracias a estos datos podemos conformar una estructura de datos que represente los atributos de un objeto así como el grado de positividad sobre el mismo. Luego, con información a priori de los intereses de los consumidores se puede crear una función que diga que tan relevantes pueden ser dichos intereses con respectos a las ideas extraídas de los comentarios, ofreciendo así un ranking altamente personificado por cliente, lo cual mejora la experiencia del usuario.

Ahora observemos los siguientes ejemplos

"El hotel, la playa y el servicio fueron excelente, pésimo y normal respectivamente"

"Cuando era pequeño fui a un hotel 5 estrellas, era maravilloso, cada día me sentía súper bien, el restaurante y las personas que trabajaban ahí eran muy buenas, la comida era maravillosa, la playa sensacional. Escuchar esta canción de Billie Eilish me hizo sentir todo lo contrario."

En el caso de la primera oración es obvio para un ser humano la categorización de cada atributo sin embargo dada la naturaleza simple de un algoritmo determinista no es posible realizar inferencia y como resultado se obtienen datos erróneos acerca de los atributos.

El caso de la segunda oración vemos un ejemplo más complejo. Aparentemente parecía un comentario sobre un hotel pero en el remate del parrafo se observa que en realidad era una crítica a sobre un tema musical. En este caso incluso aunque existen muchas palabras en el comentario que podríamos considerar como positivas, en general el mensaje es bastante negativo, por lo que una vez más se obtendrían resultados no consecuentes con la realidad.

Podemos obeservar que este algoritmo es bastante eficiente con comentarios que sean directos; sin embargo, es incapaz de discernir el ruido de la información relevante, tiene graves problemas para realizar inferencias, y en general, dada la riqueza expresiva de los lenguajes humanos, no es viable para la extracción de conceptos complejos que se podrían dar.

2.2. Algoritmos con enfoques probabilísticos y estadísticos

Existen diversas formas de abordar los desafíos del procesamiento del lenguaje natural; sin embargo, hasta la fecha no ha surgido ningún algoritmo determinista que pueda resolver estos problemas al nivel de un ser humano. Esto se debe a la incapacidad de las computadoras para realizar inferencias sin al menos una base de conocimiento a la que consultar, o sin otros modelos estadísticos que permitan generar respuestas adecuadas.

Se han desarrollado enfoques basados en aprendizaje automático que, tras un exhaustivo entrenamiento con grandes bases de datos, logran aproximar resultados significativos. No obstante, estos enfoques suelen estar limitados a secciones

específicas del conocimiento según el conjunto de datos utilizado. Aunque se han logrado mejoras en el análisis de sentimientos, no se obtiene un rendimiento adecuado para cualquier tipo de comentario en un dominio particular. Por lo tanto, es necesario contar con herramientas más robustas que sean capaces de adaptarse a diversos dominios y extraer adecuadamente los atributos relevantes. [7]

Se consideró utilizar otros métodos deterministas como:

- **Regresión lineal:** Un modelo que asume una relación lineal entre las variables dependientes e independientes y busca minimizar la diferencia entre los valores observados y predichos.
- **Árboles de decisión:** Modelos que dividen los datos en ramas basadas en ciertas condiciones, creando una estructura en forma de árbol que se puede utilizar para tareas de clasificación y regresión.
- **Máquinas de vectores de soporte (SVM):** Algoritmos capaces de encontrar el hiperplano que mejor separa las diferentes clases de sentimientos (o cualquier otro dato) de un dominio específico para discernir del ruido. [1]

- **Naive Bayes:** Un clasificador probabilístico simple pero efectivo para determinar si un texto expresa un sentimiento positivo, negativo o neutro.
- **Redes Neuronales:**
  - **Redes Neuronales Recurrentes (RNN):** Excelentes para capturar la secuencia de palabras en un texto y modelar dependencias a largo plazo. [2]
  - **Redes Neuronales Convolucionales (CNN):** Capturan características locales en el texto, como n-gramas, y son eficientes para tareas de clasificación.
  - **Transformers:** Modelos de última generación que han revolucionado el campo del PLN, como BERT y GPT, debido a su capacidad de capturar relaciones contextuales entre palabras a gran escala.

El uso de estas herramientas se ha caracterizado por el entrenamiento en un conjunto de datos específicos que responden a una intención de los autores de proponer un modelo que resuelva un problema específico de la industria. Estos modelos tienen en su mayoría la deventaja que al ser testeados con datos fuera de su 'scope' de conocimiento disminuyen la calidad de sus respuestas. Por lo que su uso para la problemática propuesta no conllevaría a resultados satisfactorios. [6] [7]

3. Implementación Final

Por las razones planteadas previamente se decidió optar por un enfoque distinto y utilizar Grandes Modelos del Lenguaje, debido a la complejidad inherente del lenguaje natural y la necesidad de capturar las sutilezas semánticas y contextuales que van más allá de las capacidades de los modelos deterministas tradicionales. Además, los GML han demostrado una gran capacidad para

aprender patrones complejos a partir de grandes cantidades de datos, lo que los hace ideales para tareas de procesamiento del lenguaje natural. Estos modelos también pueden generar texto coherente y contextualmente relevante, lo cual es fundamental para tareas específicas como la resolución de la inferencia y separación de ruido. Por último, los GML son altamente versátiles y pueden ser adaptados a una amplia gama de tareas distintos ambitos solo con la proporción de un contexto adecuado.

Otro punto importante es que es capaz de responder en una sola consulta la extracción de atributos con una adecuada puntuación de positividad lo que garantiza un mejor desempeño en contraste con los métodos anteriores.

Por estas razones se decidió utilizar el modelo de lenguaje Gemini 1.5 Flash, preajustado para el análisis semántico y con un 'prompt'para recibir las respuestas en formato JSON. Luego, una vez analizada la respuesta, con sus respectivos valores de positividad, se ponderan estos valores en conjunto con otras métricas de características de los textos.

3.1. Desventaja del modelo propuesto y de nuestra investigación

Aunque el uso de un GML resultó en una gran ventaja por la simplicidad de la implementación, su uso todavía tiene varios desafíos. El costo computacional asociado a múltiples solicitudes y el consumo excesivo de tokens, especialmente en bases de datos voluminosas, pueden resultar prohibitivos. Si bien la reducción del corpus puede mitigar estos costos, podría comprometer la calidad y la generalización del modelo. Además, la dependencia contextual y el uso de jerga especializada pueden limitar su desempeño en casos muy específicos. Los GML aprenden de los datos con los que son entrenados, por lo que pueden perpetuar los sesgos presentes en esos datos de no ser corregidos. Es difícil entender cómo los GML llegan a sus conclusiones, lo que dificulta la depuración y la identificación de errores.

3.2. Otras métricas que se utilizaron

La evaluación de la complejidad sintáctica y la diversidad léxica se llevó a cabo mediante métricas personalizadas. La complejidad sintáctica se calculó a partir de una combinación ponderada de cuatro factores: número de cláusulas principales, cláusulas secundarias, frases preposicionales y modificadores adjetivales. Los pesos asignados a cada factor fueron determinados empíricamente. Sin embargo, se sugiere explorar la aplicación de técnicas de regresión para optimizar estos pesos. Por otro lado, la diversidad léxica se cuantificó utilizando la entropía de Shannon. Esta métrica, aplicada a un corpus lematizado y tokenizado, mide la distribución de las palabras en el texto. La entropia de Shannon penaliza la alta frecuencia de términos, favoreciendo una mayor diversidad léxica. Aunque estos indicadores ofrecen una visión inicial de la complejidad lingüística del texto, se reconoce la necesidad de explorar métricas adicionales y modelos más sofisticados para una evaluación exhaustiva.

4. Conclusiones

El lenguaje humano presenta una gran complejidad, con múltiples significados, contexto y ambigüedades. Esto dificulta la creación de algoritmos deterministas que puedan capturar la riqueza y sutileza del lenguaje. Los métodos tradicionales, como los algoritmos deterministas, TF-IDF, Bag-of-Words, etc, tienen dificultades para manejar la ambigüedad y la complejidad del lenguaje natural. A menudo, se quedan cortos en tareas que requieren inferencia y comprensión contextual. Los GML han demostrado ser herramientas poderosas para el procesamiento del lenguaje natural. Esto los convierte en una opción atractiva para tareas como la extracción de atributos y el análisis de sentimientos. La implementación propone un enfoque híbrido para no dejar todo el procesamiento de los datos al GML. Los modelos deterministas pueden proporcionar una base sólida para tareas específicas, mientras que los GML pueden agregar una capa de sofisticación y flexibilidad.

5. Recomendaciones

Existen numerosos GML disponibles, cada uno con sus propias fortalezas y debilidades. Por lo que se recomienda evaluar diferentes modelos y seleccionar el que mejor se adapte a las necesidades específicas de la tarea. Otra opción viable sería la preparación mediante el "Ajuste Fino"de un modelo específico para la extracción de atributos y clasificación. Para ello sería necesario garantizar la calidad de los datos de entrenamiento, por lo que, se recomienda crear conjuntos de datos que sean representativos del dominio y que contengan una gran variedad de ejemplos. Esto además sería positivo para la realización de una rigurosa métrica para evaluar el desempeño de los modelos.

En el caso del uso de métricas adicionales se mencionaron el análisis de la diversidad léxica y la complejidad sintáctica. Además, se pudiera introducir una forma de estimar la coherencia temática del texto haciendo uso de, por ejemplo, "Latent Dirichlet Allocation(LDA)". De hecho, este enfoque fue introducido en nuestro modelo; pero por razones de tiempo y complejidad tuvo que ser lastimosamente abandonado. En cualquier caso permanece como una posible vía para mejorar nuestro sistema, pues la coherencia temática aportaría una nueva dimensión (de mucho peso además) al análisis de la estructura del texto.

Referencias

[1] Farman Ali, Kyung-Sup Kwak y Yong-Gi Kim. "Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification". en. En: *Applied Soft Computing* 47 (oct. de 2016), págs. 235-250. ISSN: 15684946. DOI: 10.1016/j.asoc.2016.06.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1568494616302691 (visitado 02-09-2024).

[2] Madhumita Guha Majumder, Sangita Dutta Gupta y Justin Paul. "Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis". en. En: *Journal of Business Research* 150 (nov. de 2022), págs. 147-164. ISSN: 01482963. DOI: 10.1016/j.jbusres.2022.06.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0148296322005446 (visitado 02-09-2024).

[3] *Knowledge Representation and Reasoning*. en. Elsevier, 2004. ISBN: 978-1-55860-932-7. DOI: 10.1016/B978-1-55860-932-7.X5083-3. URL: https://linkinghub.elsevier.com/retrieve/pii/B9781558609327X50833 (visitado 02-09-2024).

[4] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to information retrieval*. eng. Reprinted. Cambridge: Cambridge Univ. Press, 2009. ISBN: 978-0-521-86571-5.

[5] Bhargav Srinivasa-Desikan. *Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras*. eng. OCLC: 1046682463. Birmingham, UK: Packt Publishing, 2018. ISBN: 978-1-78883-703-3.

[6] G. Vinodhini y R.M. Chandrasekaran. "A comparative performance evaluation of neural network based approach for sentiment classification of online reviews". en. En: *Journal of King Saud University - Computer and Information Sciences* 28.1 (ene. de 2016), págs. 2-12. ISSN: 13191578. DOI: 10.1016/j.jksuci.2014.03.024. URL: https://linkinghub.elsevier.com/retrieve/pii/S1319157815001020 (visitado 02-09-2024).

[7] Lili Zheng. "The classification of online consumer reviews: A systematic literature review and integrative framework". en. En: *Journal of Business Research* 135 (oct. de 2021), págs. 226-251. ISSN: 01482963. DOI: 10.1016/j.jbusres.2021.06.038. URL: https://linkinghub.elsevier.com/retrieve/pii/S0148296321004458 (visitado 02-09-2024).

Bibliografía

[3] *Knowledge Representation and Reasoning*. en. Elsevier, 2004. ISBN: 978-1-55860-932-7. DOI: 10.1016/B978-1-55860-932-7.X5083-3. URL: https://linkinghub.elsevier.com/retrieve/pii/B9781558609327X50833 (visitado 02-09-2024).

[4] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to information retrieval*. eng. Reprinted. Cambridge: Cambridge Univ. Press, 2009. ISBN: 978-0-521-86571-5.

[5] Bhargav Srinivasa-Desikan. *Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras*. eng. OCLC: 1046682463. Birmingham, UK: Packt Publishing, 2018. ISBN: 978-1-78883-703-3.