Dr. Mario Graff Guerrero

Doctor en Ciencias Computacionales (PhD in Computer Science)

Investigadora o Investigador Nacional, Nivel 2 (SNII-2)
(2022-2026)

Datos de contacto:

mario.graff@infotec.mx
(55) 5624 2800 ext. 6315

Análisis de la Competitividad de Representaciones basadas en Bolsa de Palabras en Tareas de Procesamiento de Lenguaje Natural

Datos Generales

Resumen ejecutivo

En la actualidad las tareas de procesamiento de lenguaje natural (PLN) están siendo atacadas con representaciones obtenidas mediante aprendizaje profundo, dejando de lado técnicas tradicionales como las representaciones basadas en bolsa de palabras. Aunque el rendimiento de soluciones basadas en aprendizaje profundo ha sido superior en diversas tareas, la creación de estos modelos es computacionalmente intensa además de que se requieren de una gran cantidad de datos. Por otro lado las bolsas de palabras requiere una cantidad menor de recursos tanto en datos como en cómputo. Este proyecto estudiará la competitividad del uso de bolsa de palabras principalmente en tareas de PLN que se puede plantear como problemas de categorización de texto, ejemplos de estos son el identificar la polaridad de un texto, la emoción o el género de la persona que lo escribió.

Línea General y Aplicación del Conocimiento (LGAC)

10. Inteligencia computacional en la Ciencia de Datos

Palabras clave

Categorización de texto; Bolsa de Palabras; Representaciones de Texto.

Objetivo General

Analizar la competitividad de representaciones basadas en bolsa de palabras en tareas de procesamiento de lenguaje natural.

Objetivos especificos

• Comparar el rendimiento de bolsas de palabras y métodos de aprendizaje profundo en tareas de procesamiento de lenguaje natural. • Proponer, desarrollar y analizar el uso de medidas de rendimiento como función objetivo en la estimación de parámetros. • Proponer un algoritmo competitivo basado en bolsa de palabras aplicable a las diferentes tareas de PLN analizadas. • Analizar la explicabilidad de representaciones basadas en bolsa de palabras.

Datos del proyecto

Descripción

Varias tareas de procesamiento de lenguaje natural se pueden plantear como problemas de aprendizaje supervisado, es decir, problemas de clasificación o regresión. Para poder utilizar la gran mayoría de algoritmos de aprendizaje supervisado es necesario representar el texto como un vector. Uno de los procedimientos más utilizados y más simples para representar un texto en un espacio vectorial, es utilizar una bolsa de palabras. Para crear una bolsa se palabras se requiere de haber definido un vocabulario, este puede ser de palabras, secuencias de caracteres (q-gramas de caracteres), n-gramas de palabras, entre otras partes. El vocabulario se define mediante un conjunto de textos. A cada elemento del vocabulario se le asigna un identificador, este identificador corresponde a la componente del espacio vectorial. Un texto es simplemente un valor diferente de cero en todos los elementos del texto que se encuentran en el vocabulario, si un elemento de texto no se encuentra en el vocabulario se descarta. Un algoritmo que se utiliza para asignarle el peso a las componentes es el inverso de la frecuencia del término multiplicado por la frecuencia del término (TFIDF, por sus siglas en inglés). Las bolsas de palabras pueden servir para crear representaciones más ricas en contenido, en este proyecto estas representación serán referidas como representaciones densas. Una metodología podría ser el representar un texto en un espacio vectorial donde cada componente este asociado a la presencia o ausencia de una palabra, característica o emoji. Por ejemplo, asumiendo que se tienen M conjuntos de datos de datos, donde cada conjunto corresponde a un problema de clasificación de texto binario. Para cada conjunto se crear un clasificador, utilizando estos M clasificadores de texto, se puede representar un texto por la concatenación de los valores de la función de decisión de estos M clasificadores. Entonces cada componente nos dice la clase del problema de clasificación binaria. Un procedimiento para generar M problemas de clasificación de texto es utilizar aprendizaje autosupervisado. En esta técnica se inicia con un conjunto de datos no etiquetado y se termina con un conjunto de datos etiquetados, donde la etiqueta corresponde a un conjunto de reglas aplicadas a los datos. Por ejemplo, estas reglas podrían ser el buscar en el texto un emoji particular, quitar el emoji del texto y asociar aquellos textos donde se encontró el emoji buscando a la clase positiva y aquellos donde no se encontró a la clase negativa. Este procedimiento se puede realizar de la misma manera para palabras, bi- gramas o cualquier otra secuencia de caracteres. El texto puede ser representado en un espacio vectorial utilizando una bolsa de palabras y las representaciones densas, estas dos tipos de representaciones se combinarán con diferentes algoritmos de aprendizaje supervisado así como el uso simultáneo de las mismas para resolver un problema de clasificación de texto. Este proyecto utiliza esta metodología para investigar la competitividad de bolsas de palabras en problemas de clasificación de texto. Para esto se utilizará esta metodología para resolver problemas de clasificación de texto presentados en diferentes competencias internacionales. Se decide el utilizar competencias internacionales para tener un comparación con soluciones realizadas por otros grupos de investigación.

Antecedentes del problema a resolver

Los antecedentes de este proyecto se remontan al artículo EvoMSA: A multilingual evolutionary approach for sentiment analysis, realizado por Dra. Moctezuma, Dr. Téllez, Dr. Miranda-Jiménez y el responsable de este proyecto. En este artículo se propone el generar un clasificador de texto mediante la unión de diferentes representaciones de texto. Estas representaciones han sido extendidas utilizando aprendizaje autosupervisado; donde la tarea es predecir la presencia de una palabra o emoji. De esta manera se tienen representaciones densas en 18 lenguajes diferentes. Para cada lenguaje se tiene un número diferente de componentes, por ejemplo para español se tienen 2,048 palabras y 567 emojis, es decir, se generaron 2,048 problemas donde la tarea es predecir la presencia de la palabra analizada y 567 problemas donde se predice la presencia de un emoji. Estas representaciones son la base para el proyecto propuesto. Es decir, lo clasificadores de texto estarán basado en bolsas de palabras y las representaciones densas creadas con aprendizaje auto-supervisado. En la realización del artículo EvoMSA y los siguientes años se ha participado en diferentes competencias de categorización de texto, en algunas competencias se ha tenido un desempeño satisfactorio, sin embargo, no se ha logrado tener un sistema consistente para poder analizar la competitividad de las bolsa de palabras. Es decir, la dinámica de la competencia lleva a tomar decisiones que no están lo suficientemente estudiadas y por azar estas decisiones pueden resultar en un buen rendimiento que solamente es aplicable a la competencia analizada. En este proyecto se tendrá una metodología estable que permita observar el tipo de problemas de clasificación de texto donde las bolsas de palabras son competitivas.

Justificación y pertinencia

Conocer la competitividad de las bolsas de palabras en problemas de clasificación de texto es importante desde varios aspectos. En primer lugar, las bolsas de palabras son procedimientos que se pueden implementar eficientemente; para su construcción se requiere de una cantidad de datos mucho menor que las soluciones basadas en aprendizaje profundo. Complementando lo anterior, una vez entrenado el algoritmo este requiere de pocos recursos computacionales para su funcionamiento, por lo cual se pueden aplicar a una gran cantidad de datos sin la necesidad de contar con equipo especializado. En segundo lugar, las soluciones basadas en bolsas de palabras son explicables, se puede conocer las razones por las que se tomo una decisión, así como analizar el conjunto de datos que se utilizó para su entrenamiento.

Metas

• Desarrollar clasificadores de texto competitivos basados en bolsa de palabras • Desarrollar clasificadores explicables

Metodologías

Este proyecto utilizará una metodología tradicional en el desarrollo de algoritmos aplicados a problemas de aprendizaje supervisado. Se inicia seleccionando los problemas en los cuales se medirá el rendimiento de los algoritmos desarrollados. Habiendo decidido los problemas, se dividen estos en datos de entrenamiento, validación y prueba. Los datos de validación pueden ser obtenidos del conjunto de entrenamiento. Considerando que en este proyecto los datos a utilizar se basan en datos que han sido utilizados en competencias, entonces la división de los conjuntos de entrenamiento y prueba ya se encuentran especificadas en la misma competencia. Habiendo definido los conjuntos de entrenamiento, validación y prueba se procede a seleccionar las medidas de rendimiento que se utilizarán. En este caso las medidas que se utilizarán corresponden en su mayoría a medidas que se utilizan en problemas de clasificación como lo son, f1, macro-f1, cobertura, precisión, exactitud, entre otras. Se procede a optimizar los parámetros de los algoritmos desarrollados, así como sus hiperparámetros, utilizando el conjunto de entrenamiento y validación. Con los algoritmos entrenados, se mide el rendimiento de estos en los conjuntos seleccionados. Para esto se predicen las clases a las que pertenece cada elemento del conjunto de prueba y se calcula su rendimiento. Teniendo las mediciones del rendimiento para cada uno de los problemas y algoritmos seleccionados se procede a realizar un análisis estadístico para conocer si las diferencias en rendimiento son significativas.

Resultados esperados

Se describirán los resultados obtenidos en artículos científicos que serán enviados a revistas y los resultados parciales del proyecto serán enviados a congresos internacionales. Estos artículos científicos estarán disponibles en el repositorio arxiv permitiendo así su lectura libre. Además de los artículos científicos, los algoritmos desarrollados serán puestos a disposición del público general en la plataforma GitHub. Además del código fuente, se creara la documentación necesaria para que estos puedan ser utilizados por usuarios externos.

Cronograma de trabajo

#	Entregable(s) comprometido(s)	Fecha inicio	Fecha fin
1	Artículo de investigación en revista especializada	01/04/2024	31/12/2024
2	Impartición de docencia.	01/01/2024	31/05/2024
3	Impartición de docencia.	01/08/2024	31/12/2024
4	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
5	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
6	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
7	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
8	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
9	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
10	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
11	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
12	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
13	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
14	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
15	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	01/01/2024	31/12/2024
16	Divulgación	01/01/2024	31/12/2024
17	Divulgación	01/01/2024	31/12/2024

Bibliografía relevante

Tellez, E.S., Moctezuma, D., Miranda, S. et al. Regionalized models for Spanish language variations based on Twitter. Lang Resources & Evaluation 57, 1697–1727 (2023). https://doi.org/10.1007/s10579-023-09640-9 M. Graff, S. Miranda-Jimenez, E. S. Tellez and D. Moctezuma, "EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis [Application Notes]," in IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 76-88, Feb. 2020, doi: 10.1109/MCI.2019.2954668. Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, Elio A. Villaseñor, A case study of Spanish text transformations for twitter sentiment analysis, Expert Systems with Applications, Volume 81, 2017, Pages 457-471,