Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRAO
GRAO EN ENXEÑARÍA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

Sistema Deep Learning para el análisis de sentimientos en opiniones de productos para la ordenación de resultados de un buscador semántico

Estudante: Adrián Insua Yañez
Dirección: Carlos Gómez Rodríguez
Dirección: Sonia González Vázquez

A Coruña, novembro de 2019.

Resumen

En esta investigación abordaremos el problema conocido como Análisis de Sentimientos, enmarcado dentro del área de estudio del Procesamiento de Lenguaje Natural, también llamado NLP por sus siglas en inglés (Natural Language Processing). La tarea consiste en identificar y extraer la polaridad de un conjunto de textos que expresan opiniones de personas con el objetivo de clasificarlos.

El desarrollo de este tipo de tareas de clasificación ha adquirido gran relevancia en los últimos tiempos dada su potencial aplicación al mundo empresarial y al crecimiento exponencial de los conjuntos de datos disponibles para realizar la investigación, gracias al uso cada vez más común de las redes sociales.

En cuanto a la importancia en el ámbito empresarial, a cualquier entidad que tenga un producto en el mercado le puede resultar interesante conocer la opinión que tienen sus clientes sobre la calidad de sus productos automáticamente, pudiendo saber tanto una opinión del público general, como realizando estudios sobre zonas geográficas determinadas. De igual forma este tipo de clasificaciones se pueden aplicar sobre otros ámbitos como por ejemplo el político, analizando el nivel de descontento de la población según la polaridad resultante de un conjunto de tweets filtrados por un hashtag (etiqueta señalada con #) determinado, sin necesidad de analizar manualmente el gran volumen de datos que puede suponer este tipo de estudios.

En este trabajo se pretende investigar distintas técnicas de aprendizaje automático aplicadas al campo NLP para obtener un análisis fiable de la información subjetiva de un conjunto de textos. En este sentido abordaremos el estudio de algoritmos de aprendizaje automático clásico (Machine Learning), que servirá para establecer una línea base sobre la que intentar mejorar los resultados, para posteriormente implementar algoritmos más modernos de aprendizaje profundo (Deep Learning), con la intención de que este tipo de sistemas sean capaces de aprender a discernir la estructura de las oraciones y gracias a ello mejorar los resultados de clasificación obtenidos.

Dado que la investigación se presenta dentro de un marco profesional, se ha orientado al dominio del problema específico. En este caso se trata de un sistema que ha de clasificar un conjunto de opiniones en español sobre materiales de construcción, para posteriormente utilizar estas polaridades en un sistema de ranking que se aplicará a los resultados de un buscador semántico utilizado por usuarios expertos en el sector. Es importante que el usuario acceda rápidamente a los mejores productos de forma que mejoremos su experiencia y consigamos una mayor fidelización.

Para la validación del clasificador se ha desarrollado un sistema que además de clasificar los textos según las polaridades encontradas, devuelve un conjunto de métricas que explicaremos más adelante, y que nos permitirá comparar el funcionamiento de los distintos algoritmos.

De igual forma, aunque de manera secundaria, el trabajo abordará la implementación de la aplicación tanto en lo tocante al servidor como a la parte web. Esta aplicación tendrá una sección de comentarios en el perfil del producto para que el usuario pueda dar su opinión sobre el mismo, en esta sección se le permitirá además establecer una puntuación utilizando un sistema de "estrellas" típico que clasificará el texto en un rango de 1 a 5 siendo 1 muy negativo y 5 muy positivo. Estas clasificaciones se utilizarán en un futuro para mejorar el modelo de clasificación utilizando un corpus perteneciente al dominio del problema.

Abstract

In this investigation we will address the problem known as Sentiment Analysis, framed within the area of study of Natural Language Processing or NLP. Our task is to identify and extract the polarity over a set of texts that express opinions from people with the objective of classifying them.

The development of this kind of classification tasks has acquired great relevance in recent times given its potential application in business and the exponential growth of available data sets for research, thanks to the increasingly common use of social networks.

As for the relevance in this business area, any entity that has a product in the market may find interesting to know the opinion that their customers have about the quality of their products automatically, being able to know both an opinion of the general public, and making studies on specific geographical areas. In the same way, this type of classifications can be applied to other areas such as the political one, analyzing the level of discontent of the population according to the polarity resulting from a set of tweets filtered by a given hashtag (label marked with #), without needing a manual anlysis of the large volume of data that this type of study may entail.

This work aims to investigate different machine learning techniques applied to the NLP field to obtain a reliable analysis of the subjective information of a set of texts. In this sense we will approach the study of classic machine learning algorithms, which will serve to establish a baseline on which we will try to improve the results implementing more modern deep learning algorithms by discerning the sentence's structure.

Since the research is presented within a professional context, it has focused on mastering the specific problem. In this case it is a system that has to classify a set of opinions in Spanish on construction materials, to later use these polarities in a ranking system that will be applied

to the results of a semantic search engine used by sector experts. It is important for the user experience to get the best products quickly, so we can achive greater loyalty.

For the validation of the classifier a system has been developed that in addition to classifying the texts according to the polarities found, returns a set of metrics that we will explain later, and that will allow us to compare the different algorithms.

Similarly, although in a secondary way, the work will address the implementation of the application both in terms of the server and the web part. This application will have a comments section in the product profile so that the user can give his opinion on it, in this section he will also be allowed to establish a score using a typical " stars " system that will classify the text in a range from 1 to 5 where 1 means very negative and 5 very positive. These classifications will be used to improve the classification model using a corpus belonging to the problem domain.

Palabras clave:

- Procesamiento de lenguaje natural
- PLN
- Análisis de sentimientos
- Minería de opiniones
- Aprendizaje automático
- Aprendizaje profundo
- Análisis subjetivo

Keywords:

- Natural language Processing
- NLP
- Sentiment analysis
- Opinion mining
- Machine Learning
- Deep Learning
- Subjective analysis