Islas Canarias
Del 15 al 19 de noviembre de 2021

WEB SCRAPING PARA CARACTERÍSTICAS DE EMPRESAS

Iparraguirre García, Ander
a-iparraguirregarcia@eustat.eus

Ledesma Sereno, Matias
m-ledesma@eustat.eus

1. Introducción

Eustat, el Instituto Vasco de Estadística, consciente de la potencia de las nuevas tecnologías para el uso de internet para recolectar información y alineado con los objetivos del ESSnet Big Data, ha lanzado un proyecto ya realizado por otros países europeos. El objetivo del mismo es el uso del webscraping, text mining e inferencia estadística para recoger y procesar información de empresas y mejorar o actualizar la información registrada en el directorio de actividades económicas de EUSTAT y variables estadísticas de diferentes encuestas.

Basado en dicho proyecto y con la colaboración de la Universidad del País Vasco (EHU-UPV) se ha lanzado un proyecto para scrapear las webs de las empresas para el estudio de la presencia en redes sociales y el posible uso de dicha información para la encuesta ESIE (Encuesta de la sociedad de la información a empresas). La técnica webscraping presenta una oportunidad única para mejorar la calidad de los registros sin incrementar la carga de respuesta. Se han detectado varios ítems en la encuesta relacionados con la información objetivo del proyecto como por ejemplo "Indique si su web ofrece alguna de las siguientes prestaciones: Enlaces a los perfiles de la empresa en los medios sociales (Facebook, twitter, …)" que podrían ser eliminados.

Tabla 1. Encuesta ESIE del Eustat sobre Medios de Comunicación Sociales

B.3.5. Uso de los medios de comunicación sociales (Indique cuáles utiliza su empresa)
a) Redes Sociales (Facebook, LinkedIn, Tuenti, Google+, …)
b) Blogs de empresas o microblogs (Twitter, Blogger, …)
c) Webs para compartir contenido multimedia (Youtube, Flickr, Instagram, …)

La obtención de la presencia en redes sociales en internet es sólo una de muchas variables estadísticas que se puede obtener escrapeando información en internet. Entre otras posibles variables también se encontrarían: “Tiene página web”, “Indique cuáles son los idiomas utilizados en sus páginas web” o “Publicidad de ofertas de empleo” que se engloban en este proyecto.

2. Objetivos

Tratar de obtener una nueva fuente de datos utilizando técnicas de webscraping para sustituir o validar a los obtenidos vía encuesta. El objetivo ha sido compara los resultados de la encuesta y lo que se obtiene mediante la técnica de webscraping.

3. Metodología

Con el objetivo de testear este piloto, se ha usado la muestra de establecimientos extraídas de la población de referencia incluida en la encuesta sobre la sociedad de la información de las empresas en Euskadi en el año 2021 incluyendo sólo aquellas de las que conocemos su web y a su vez también se han eliminado aquellas URL que presentaban problemas o no estaban disponibles. De esta manera, tenemos un subconjunto de compañías con información vía web y encuesta. El número de empresas seleccionadas con este proceso es de 3715.

Centrarnos en esta muestra permite tener una visión general de manera rápida de los datos erróneos tomando como verdaderos los datos recogidos a través de la encuesta. Una vez diseñada la lista de URLs disponibles comenzamos con la detección de redes sociales usando el software desarrollado por Jacek Maslankowski[1] refinado para añadir una nueva fase de recolección de datos específica para detectar redes sociales en código JavaScript.

Finalmente se realiza una comparación de las diferentes preguntas relacionadas con la presencia de redes sociales de la encuesta ESIE, frente a la predicción basada en webscraping.

4. Resultados

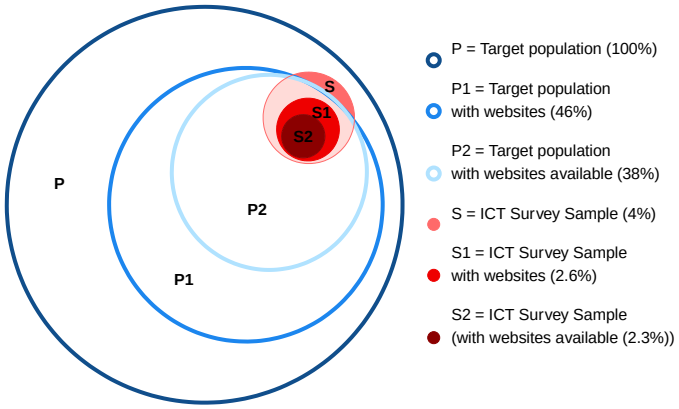
A continuación, se presentan los resultados de la configuración de la muestra y la comparación de los resultados a partir de la misma.

4.1. Muestra

En la figura 1 se puede comprobar cómo ha sido seleccionada la base de datos de entrenamiento siguiendo el proceso explicado en la parte de metodología. Del universo de compañías vascas, la base de datos se reduce a un subconjunto del 2,31% de la población. Esta figura refleja perfectamente cuales son las posibles limitaciones de la base de datos de entrenamiento, pero también muestra una sencilla manera de incrementar el tamaño muestral hasta el 38% de la población sin un alto coste computacional.

[1] https://github.com/essnetbigdata/WP2-Social-Media-Presence

Figura 1. Diagrama de Euler del proceso de selección de los datos para el entrenamiento.

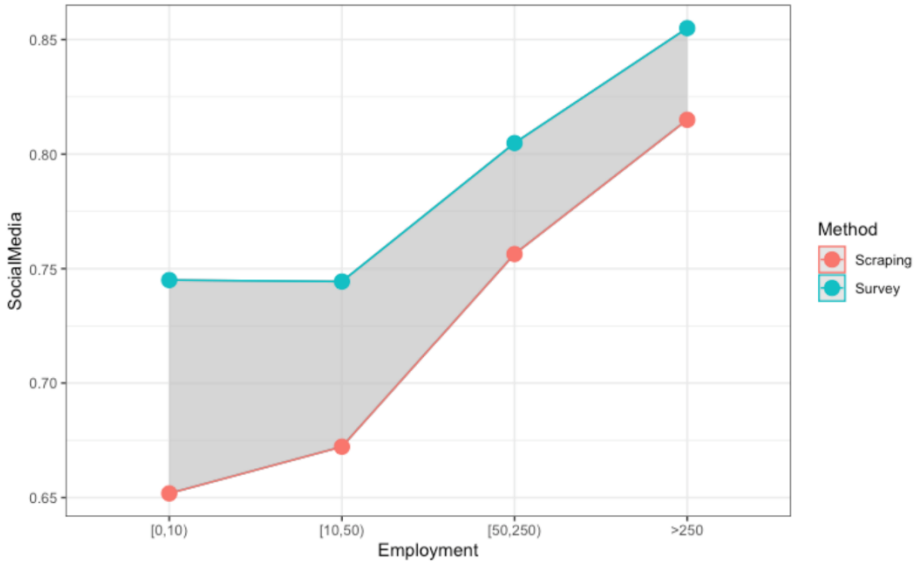


4.2. Comparación de resultados

Una vez está seleccionada la muestra se comparan los resultados de la encuesta tomados como ciertos y los resultados obtenidos de la predicción vía escrapeo. El webscraping estima que el 69% de las compañías están usando medios de comunicación sociales basados en la muestra S2 (no elevada). Este resultado se compara con el 76,33% que respondieron afirmativamente a la pregunta “Utiliza su compañía alguna red social?” (No elevado, de la encuesta ESIE 2021).

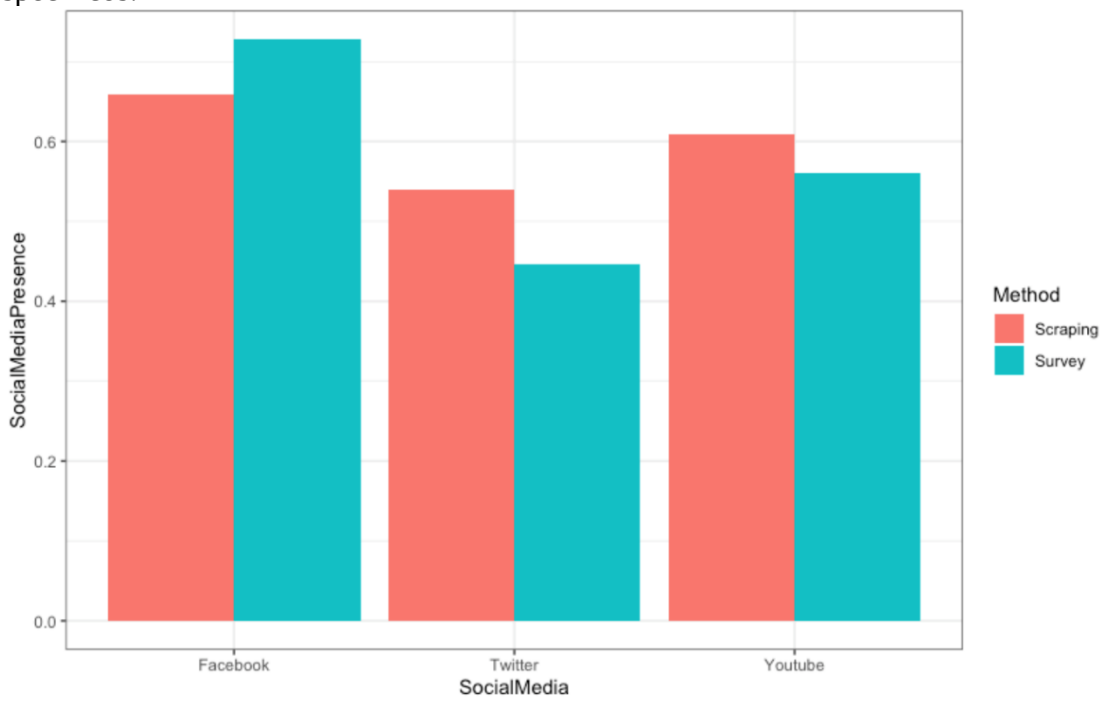
Realizando una exploración de los datos, la discrepancia entre la estimación y los resultados de la encuesta se podría deber a la falta de links de redes sociales en las páginas web de las micro y pequeñas empresas. Como podemos apreciar en la Figura 2, la brecha entre la encuesta y los resultados del escrapeo tiende a converger cuando el tamaño de la compañía se incrementa medido por el número de empleados. Estos resultados son coherentes con algunos de los resultados de otros países, que solo realizan estas estimaciones con compañías con más de 10 empleados.

Figura 2. Gráfico de puntos para reflejar como convergen la encuesta y la predicción a medida que el tamaño de las compañías se incrementa.



En relación con las estadísticas presentes en la Tabla 1 presentada anteriormente que recogen clasificaciones más específicas de medios de comunicación social, a continuación, se puede observar en la Figura 3 como el método de webscraping es razonablemente preciso en compañías de más de 10 empleados incluso para tipos concretas. En el caso de las clasificadas como redes sociales se consiguen captar el 91% de los positivos presentes en la encuesta. Por otro lado, se encuentran resultados más extraños al analizar los resultados de la presencia en micro blogs y contenidos multimedia obteniendo más presencia en los mismo por la predicción que por el método tradicional.

Figura 3. Gráfico de barras comparando resultados en medios de comunicación específicos.



Conclusión

Los resultados arrojan una perspectiva razonable para la obtención por la vía alternativa presentada en este trabajo una vez se ha refinado el software original. El tratamiento manual para la preparación de la base de datos de entrenamiento resultó costosa a nivel de consumo de tiempo sin embargo el EUSTAT ya ha lanzado un proyecto paralelo con el objetivo de solucionar dicho problema. Una vez esté completado se podrá observar unos mejores resultados aplicando esta metodología.

La predicción realizada por webscraping estima que el 69% de las compañías está usando medios de comunicación social basados en la muestra diseñada. Este dato contrasta con el resultado que se obtiene por la vía tradicional de la encuesta que arroja un 76%.

La discrepancia en los resultados se puede deber a la presencia de varios errores. En primer lugar, se ha detectado que la corrección manual de la muestra utilizada parece no ser suficientemente precisa y difícilmente escalable si se aumenta el volumen de la encuesta. Se esperan grandes resultados del proyecto piloto para solucionar dicho problema. Por otro lado, una vez se discrimina por tamaño de empresas los resultados son esperanzadores incluso si ponemos el foco en redes sociales concretas, lo que permitiría poder publicar datos más desagregados que hasta ahora no se habían publicado.

Finalmente, se está afrontando el reto de encontrar una solución para elevar los resultados de la muestra a la población objetivo. Para futuras pasos en la investigación, la solución de la presencia de falsos positivos, el aumento de la muestra refinando el algoritmo que permita acceder a más URLs y la búsqueda para detectar respuestas en la encuesta que son erróneas.