# Dirección de Desarrollo Digital

## Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

# DETECCIÓN Y ANÁLISIS DE SECTORES ECONÓMICOS CON PERCEPCIÓN DE COMPETENCIA DESLEAL

## Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Dirección de Innovación y Desarrollo Empresarial

| Sector | Lenguaje | Fuente de datos |
|---|---|---|
| Economía | Python | Twitter y comentarios en noticias del portal web de El Tiempo |

## Presentación

La Dirección de Innovación y Desarrollo Empresarial (DIDE) ha identificado la necesidad de conocer la percepción del público en general en materia de competencia desleal con respecto a los sectores económicos. Para ello, se planteó el presente proyecto en el cual se buscó identificar sectores, empresas o bienes y servicios entre otros, de la economía colombiana que presenten mayor percepción de competencia desleal principalmente en redes sociales y en portales de medios de comunicación por medio de técnicas de analítica. De tal modo, para el desarrollo de este proyecto se utilizaron técnicas de análisis de texto y se exploraron diferentes modelos no supervisados que permitan identificar los sectores de la economía colombiana que presentan mayor percepción de competencia desleal.

*The Directorate of Innovation and Business Development (DIDE) has identified the need to know the perception of the general public regarding unfair competition with respect to economic sectors. For this, the present project was proposed in which it was sought to identify sectors, companies or goods and services, among others, of the Colombian economy that present a greater perception of unfair competition, mainly in social networks and in media portals by means of techniques analytics. Thus, for the development of this project, text analysis techniques were used and different unsupervised models were explored that allow identifying the sectors of the Colombian economy that present the highest perception of unfair competition.*

## Objetivo general

Identificar en redes sociales y portales de noticias los sectores de la economía colombiana, bienes y servicios con mayor percepción de problemas de competencia desleal.

## Objetivos específicos

1. Extraer de redes sociales y del portal de noticias de El Tiempo los comentarios y opiniones relacionadas al tema con base en una lista de términos clave suministrada por la DIDE.
2. Clasificar los comentarios y opiniones relacionados a términos de competencia desleal de acuerdo con los sectores, bienes y servicios o mercados entre otros de la economía colombiana.
3. Crear una herramienta de visualización que permita analizar los resultados.

**Metodología**

Para el desarrollo de este proyecto, se abordó el problema por medio de diferentes pasos que involucran desde la adquisición de los datos hasta la presentación de resultados. En esta sección se presentarán en orden los pasos de la metodología propuesta para la elección de comentarios y la identificación de sectores, mercados o bienes y servicios entre otros que puedan identificar una percepción de competencia desleal.

Primero se realiza la descarga de información, que proviene de comentarios de redes sociales (Twitter) y de comentarios de noticias del portal de El Tiempo, luego se realiza un preprocesamiento de texto, filtros de información y finalmente se despliega la solución por medio de un aplicativo web. Las siguientes secciones describen en detalle los procesos en el desarrollo del aplicativo desde la descarga de información hasta el despliegue del aplicativo.

### *1. Descarga de información*

Para la descarga de información de texto correspondiente a comentarios relacionados a competencia desleal, se tuvieron en cuenta dos fuentes, la primero corresponde a comentarios de redes sociales (concretamente Twitter) que tengan algún tipo de relación con los temas de interés. El segundo corresponde a comentarios de portales de noticias (concretamente El Tiempo) que tengan algún tipo de relación con economía en general para luego filtrar por el tema de interés. Dentro de dicha adquisición, en ambos casos, el principal reto es el filtro de los cometarios que realmente puedan aportar al análisis, dado que una gran cantidad de información puede ser irrelevante para este análisis.
Esta descarga se establece mediante una rutina que, de manera automática, hace la descarga de comentarios relacionados y estos son guardados en un archivo separado por comas. La descarga de *tweets* se hizo de acuerdo con los términos de búsqueda enviados por la DIDE y, adicionalmente, se descargaron comentarios ajenos al tema de interés (música, videojuegos, libros, entre otros), con el fin de capturar similitudes de temas con comentarios totalmente irrelevantes para el análisis y una vez identificados, filtrarlos del total de información.

- ***Información de Twitter***

Para la descarga de información de Twitter se utilizó la librería *tweepy* de Python que permite la descarga masiva de *tweets* con base en una serie de argumentos de búsqueda. Dichos argumentos son principalmente dos. El primero es por medio del identificador de cuentas específicas y el segundo es por medio de los términos que contengan el *tweet*. Para el primer caso, la descarga de la información se hace con base en el identificador de la cuenta de T*witter* de la forma @NombreCuenta, de tal modo que se descarga la información correspondiente a una cuenta especifica. Por otra parte, el segundo argumento es por medio de un término de búsqueda especifico, de tal modo que la librería recibe un término y esta va a buscar los *tweets* que contengan dicho término, por ejemplo, si se quiere descargar *tweets* que contengan el término "televisor", la librería va a buscar todos los *tweets* que contengan este término independientemente del contexto en el cual este se mencione. En ambos casos, la descarga se puede limitar de acuerdo con el idioma y zona, de tal modo que así se puede evitar la descarga innecesaria de *tweets* en países o idiomas que no son relevantes para el análisis.

Para este proyecto, la DIDE suministró una serie de términos relacionados a competencia desleal tales como colusión, cartelización, entre otros. De tal modo que con base en estos términos clave, se realiza la descarga de *tweets* desde el momento en que se ejecuta hasta que no encuentre más *tweets,* es decir que se hace la descarga desde un momento del tiempo específico (aquellos tweets posteriores a determinada fecha) hasta el momento actual. Por tal motivo, la descarga se hace desde el presente hasta el último valor que encuentre en el pasado. Esto permite descargar una gran cantidad de información relacionada a este tema que permita dar mayor diversidad en la información y mejorar los resultados.

Esta información es almacenada en una tabla estructurada en la cual se tiene información de fecha, *tweet*, término de búsqueda entre otros.

- ***Información de portales de noticias (El Tiempo)***

Por su parte, la descarga de información de portales de noticias se hizo por medio de *web scraping* que consiste en entrar a las páginas web por medio de un programa desarrollado en *Python* e ir recorriendo cada uno de los enlaces de noticias que puedan ser de interés para el análisis. Se eligió el portal de El Tiempo dado que es la única que permitía la descarga de comentarios de noticias sin alguna restricción, dado que los otros portales de plano no tenían sección de comentarios o la página tenía un diseño que limitaba esta descarga.

Primero se hizo la descarga de los enlaces de noticias que puedan ser de interés, por tal motivo se accedió al portal de noticias por medio de su versión en RSS que es un estándar en los portales de noticias de todo el mundo, lo cual facilita la comunicación de los portales con una máquina. Estas noticias se encuentran en formato XML y contienen los enlaces de las noticias en el periodo de tiempo en que se hace la consulta. Se tuvieron en cuenta noticias de las secciones de Economía, Sectores Económicos y Empresas.

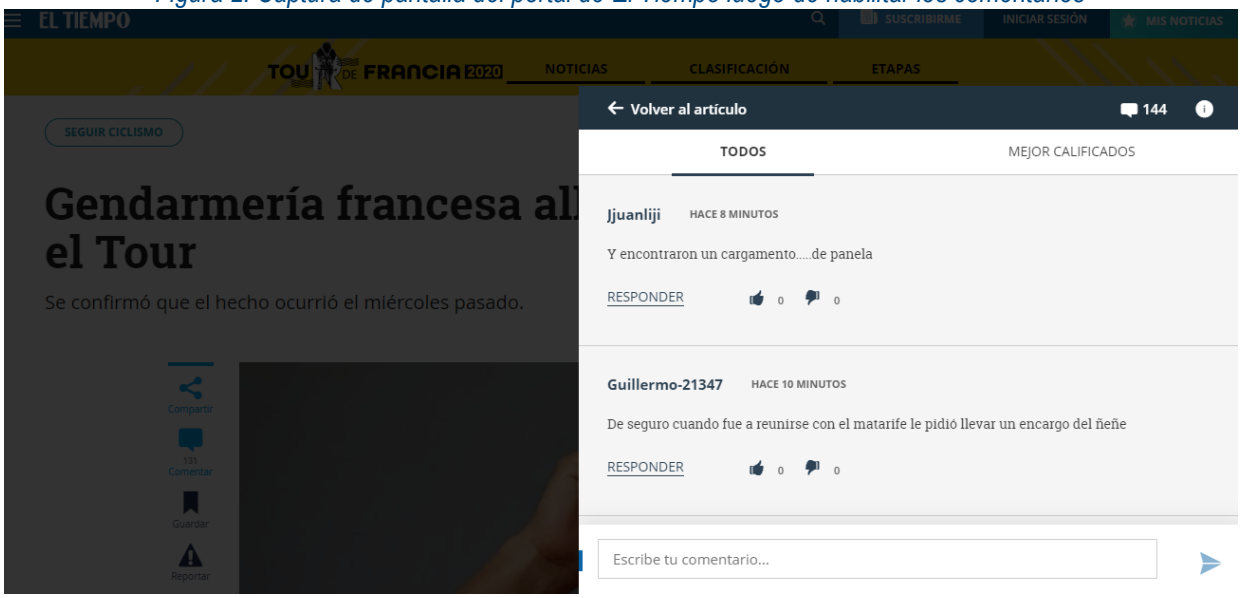
En el desarrollo del programa en Python, se utilizó la librería *Selenium* que permite entrar directamente a la página web por medio de un navegador y controlar el navegador de manera automática. Esto permite realizar clics automáticos y copiar la información necesaria. Se utilizó esta herramienta dado que la página de El Tiempo tiene un contenido que se habilita una vez se haga clic en el botón.

La Figura 1 muestra una captura de una noticia del El Tiempo, la cual muestra en la parte izquierda de la página el botón de comentarios. Este botón habilita la sección de comentarios en la Figura 2, la cual permite revisar comentarios o realizar uno nuevo. Esta sección se habilita únicamente si se hace clic en el botón anterior. Por tal motivo, es necesario interactuar de esta forma por medio de *Selenium* con cada una de las páginas web.

*Figura 1: Captura de pantalla del portal de El Tiempo de una noticia*


*Fuente: elaboración propia*

*Figura 2: Captura de pantalla del portal de El Tiempo luego de habilitar los comentarios*




*Fuente: elaboración propia*

### *2. Identificación de Stopwords*

Un tipo de *stopwords* son una serie de palabras que son comunes dentro del idioma que (entre otras cosas) pueden servir para darle un sentido gramatical a las oraciones. Estas palabras se caracterizan por darle un sentido a la oración. Sin embargo, para este análisis, no aportan información. Un ejemplo de ello son conectores, artículos, preposiciones, etc., todas palabreas que tienen importancia dentro de la oración dentro de su sintaxis, pero no hacen parte de las palabras clave de estas. Adicional a esto, estas palabras son comunes en las oraciones por lo que, al realizar un análisis de palabras, estas palabras van a tener una importancia alta dado que su ocurrencia es transversal a todas las oraciones.

Primero se utilizaron listas predefinidas de *stopwords* para el idioma español que se encuentran en internet. Sin embargo, existen palabras muy específicas que pueden considerarse como *stopword* dado que es común pero no aporta más al análisis, estas últimas dependen del objetivo del análisis y se van a tomar basados en ello. Un ejemplo de estas palabras son nombres de países, departamentos, ciudades, nombres de páginas web y en general palabras que no aportan al análisis.

Para encontrar estas palabras se hace por medio de las frecuencias de las palabras, las palabras con mayor frecuencia suelen ser este tipo de palabras, sin embargo, esta selección se hace con intervención humana, dado que, si se eliminan las n palabras con mayor frecuencia, podría estar eliminando una gran cantidad de *stopwords* pero también se podrían estar eliminando palabras que puedan aportar al análisis. Esta lista de palabras se adquiere una vez y se guardan en una lista para ser utilizadas cada vez que se necesite.

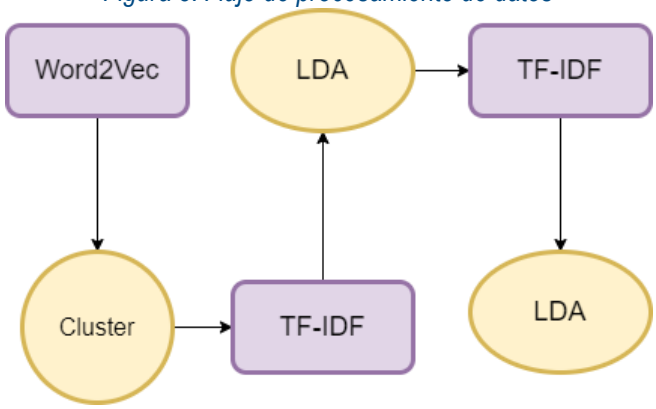
Una vez se identificaron las *stopwrods*, se realizó una limpieza de texto en las cuales incluye la eliminación de estas palabras, estandarización de palabras, eliminación de caracteres especiales, entre otros. Esto con el fin de estandarizar el texto y facilitar la implementación de los modelos.

### *3. Implementación de modelos no supervisados y vectorizaciones*

Los modelos utilizados para el procesamiento de texto son 1) Modelo Word2Vec, 2) *clustering* con seis clases, 3) primera vectorización TF-IDF, 4) LDA con 10 componentes, 5) segunda vectorización TF-IDF, 6) LDA con 3 componentes siendo este último la cantidad de *clusters* finales para mostrar en el reporte. La Figura 3 muestra el flujo

por el cual se planteó el proceso desde que ingresan los datos hasta llegar a un resultado, los modelos de *Word2Vec* y *TF-IDF* (morado) son modelos de vectorización de texto, es decir, modelos que permiten representar el texto dentro de un espacio numérico y los modelos de *Cluster* y *LDA* (Naranja) son modelos de agrupación de características que permiten asignar categorías a la información de acuerdo a similitudes entre registros.

*Figura 3: Flujo de procesamiento de datos*




*Fuente: elaboración propia*

### *3.1. Vectorización de comentaros con Word2Vec*

Con base en la información recolectada, el objetivo de la vectorización con base en el modelo de *Word2Vec* es capturar el sentido de los comentarios. De tal modo que, representar en un espacio numérico todos los comentarios y lograr capturar, dentro de ese espacio, una cercanía entre aquellos comentarios que compartan un sentido mutuo (distinguir entre aquellos comentarios que hablen de música frente a comentarios que hablen de economía). Así, esta vectorización retorna una matriz de *n* filas correspondiente a la cantidad de documentos y *k* columnas correspondiente a una cantidad definida de atributos que definen cada documento.

### *3.2. Cluster K-Medias*

Los modelos de clustering hacen parte de la rama del aprendizaje no supervisado, que su objetivo en encontrar las relaciones intrínsecas y los patrones contenidos en una matriz de información en la cual, las instancias o vectores (representadas como las filas de la matriz) y los atributos o dimensiones (representados en sus columnas).
De acuerdo con la información representada en un espacio numérico con la vectorización *Word2Vec*, se busca filtrar aquellos comentarios que no tienen ninguna relación con el tema de interés. Por esta razón, la descarga de comentarios cuyo término de búsqueda sea ajeno al tema de interés, permite relacionar aquellos comentarios que se descargan con los términos de búsqueda de interés, pero que su contenido no es relevante. Un ejemplo son los *tweets* que contienen la palabra “Cartel” pero que estos hablan de carteles del narcotráfico, el objetivo es establecer esos comentarios y alejarlos de aquellos comentarios que tienen también el término de “Cartel” pero que hablan sobre algún tema de economía.

### *3.3. 1ra Vectorización TF-IDF*

Una vez filtrado los comentarios por *cluster*, se hace una nueva vectorización con base en la matriz TF-IDF, que establece una importancia relativa de los términos y su ocurrencia en cada uno de los documentos (en este caso comentarios). De tal modo que esta vectorización se utiliza para luego capturar temas innecesarios o alejados al objetivo de estudio, que no se lograron capturar dentro de la clusterización anterior.

### *3.4. 1er modelo LDA*

El modelo LDA *(Latent Dirichlet allocation)* hace parte del grupo de modelos de aprendizaje no supervisado que permite identificar grupos no observados a partir de las mismas observaciones. Este modelo tiene un amplio desarrollo en procesamiento de lenguaje natural en el cual postula que cada documento es una mezcla de un determinado número de temas y que la presencia de cada palabra en cada documento es atribuible a cada tema.
Con base en lo anterior y basados en la vectorización TF-IDF anterior, se calcula un modelo de detección de temas LDA con 10 componentes para identificar y segmentar aquellos temas que puedan estar alrededor de un mismo tema, pero no necesariamente hacen parte del objetivo del análisis. Por ejemplo, temas de política exterior o relaciones internaciones.

### *3.5. 2da Vectorización TF-IDF*

Se realiza una última vectorización TF-IDF dado que, al tener una menor cantidad de comentarios con respecto a la primera vectorización TF-IDF, la importancia relativa de las palabras se va a ver afectada. Por tal motivo, lo que se busca con esta vectorización es que se logre capturar los temas latentes dentro del subgrupo de comentarios filtrados.

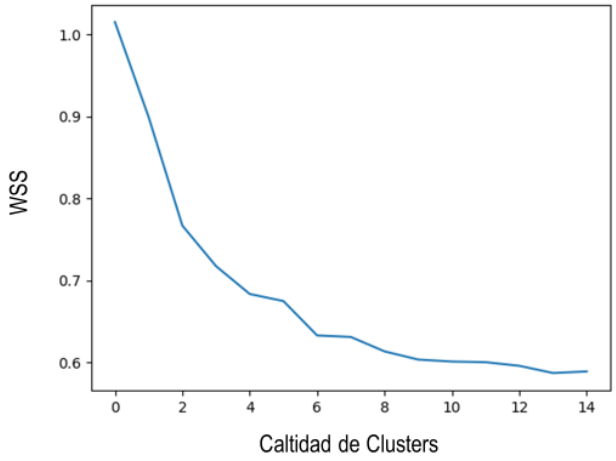
### *3.6. 2do modelo LDA*

Este último modelo tiene como objetivo capturar los temas latentes dentro del subgrupo de comentarios filtrados en las etapas anteriores. De tal modo que los resultados de este modelo son los resultados presentados en la visualización, que permite a la DIDE entregar mirar y asociar los resultados obtenidos para la identificación de sectores, bienes y servicios o empresas e la economía con percepción de competencia desleal. Los resultados de este último modelo son presentados dentro del aplicativo final, en el cual el usuario tiene la posibilidad de definir y visualizar la cantidad de temas (o grupos) que desea desde 2 a 9 grupos.

## Resultados

Dentro de la metodología anterior mencionada, los resultados intermedios del análisis de *clusterización* son presentados a continuación. A partir de la primera vectorización (*Word2Vec),* se hizo un cálculo de valores de la suma de errores al cuadrado dentro de cada cluster *(Within-cluster-sum of squared Errors (WSS))* para diferentes valores de *cluster k.* Esta métrica se calcula como la distancia al cuadrado media de cada punto con respecto al centroide del *cluster* en el cual pertenece cada categoría. Esta métrica siempre va a tender a disminuir con forme aumentan la cantidad de *clusters*. De tal modo que se busca la cantidad de *cluster* óptima como el momento en el cual se identifica un punto de inflexión en la caída del *WSS,* en el cual el cambio marginal del error de un *cluster* a otro se acerca a cero. La Figura 4 muestra los resultados de la métrica hasta un total de 15 *clusters*. En ella se muestra que existe un punto de inflexión bastante sutil en un total de 6 *clusters*. De tal modo que, para este primero modelo, se calculó un modelo *k-medias* con 6 *clusters*.

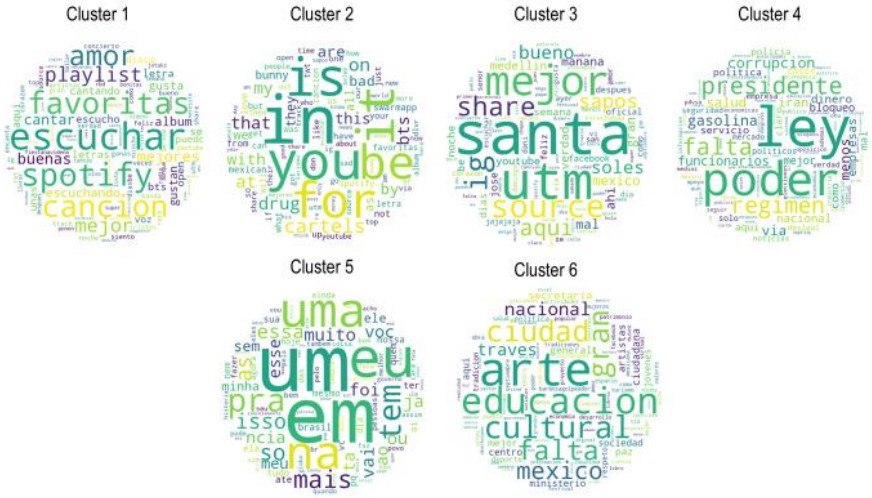
*Figura 4: Resultados WSS para diferentes cantidades de cluster*




*Fuente: elaboración propia*

Por su parte, los resultados de las nubes de palabras para cada *cluster* de información son mostrados en la Figura 5. En ella se muestra que los *clusters* logran capturar 1) comentarios en otros idiomas como inglés y portugués y 2) temas relacionados a otros campos que no son de interés. El *cluster* que pareciera tener mayor relación con el objetivo de este análisis, es el *cluster* 4, por tal motivo, se utilizará este *cluster* para seguir con el análisis dentro de subtemas identificados con LDA.

*Figura 5: Nubes de palabras para los 6 clusters*



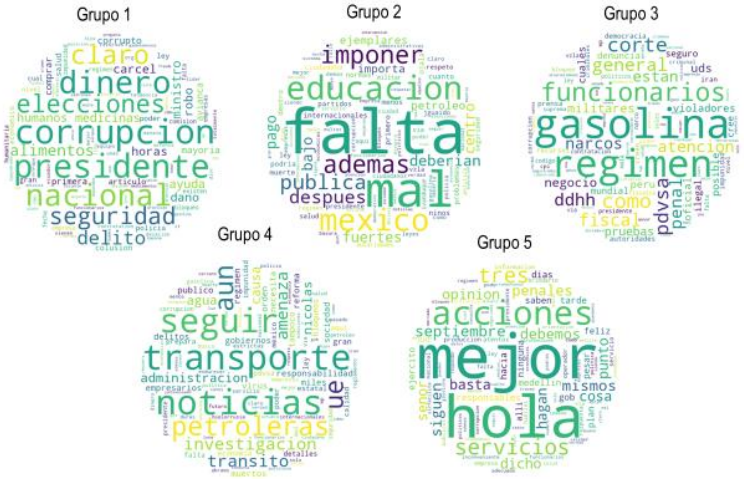

*Fuente: elaboración propia*

Posteriormente, con base en el *cluster* identificado en la etapa anterior, se dividió este *cluster* en 10 temas con base en el algoritmo LDA, de tal modo que se puedan identificar los subtemas latentes dentro del *cluster* en cuestión (se eligió un número elevado con el fin de hacer una división más desagregada de los tweets y así tener un filtro más detallado, se puede caer en el riesgo de estar dividiendo dos temas que tengan un alto grado de similitud pero aquellos temas que tengan más disimilitud, se podrán identificar con mayor detalle).

La Figura 6 y Figura 7 muestran la subdivisión de temas latentes de los primeros 5 grupos y los siguientes 5 grupos respectivamente. La definición de temas de interés se hace con base en los términos que se esperarían tener alguna relación directa o indirectamente y desechar aquellos grupos que tengan relación con otros temas sin interés. Por ejemplo, el grupo 6 puede estar mostrando información relacionada a temas internacionales dado que menciona

términos tales como “ONU” e “Irán”. Por su parte el grupo 7 tiene términos tales como ”Irán” y “Narco”. Otros grupos más ambiguos como el 5 y el 2 que tiene el término “México” con una frecuencia alta también son desechados. De tal modo que, al eliminar los comentarios con estos temas, resulta una cantidad de comentarios de alrededor del 17% del total original. Esta cantidad final, es la utilizada para mostrar en el reporte a continuación presentado.

*Figura 6: Resultados de grupos del primero modelo LDA (1-5)*




*Fuente: elaboración propia*

*Figura 7: Resultados de grupos del primero modelo LDA (6-10)*


*Fuente: elaboración propia*

El usuario tiene la posibilidad de elegir en cuántos grupos desea dividir la información. Para cada uno de estos grupos se realizaron sus respectivas nubes de palabras, gráfico de frecuencias y gráfico de co-ocurrencia. Con esta desagregación por grupos, se busca identificar la información más relevante de los temas en los cuales más se habal para cada grupo con base en la frecuencia de las palabras más usadas. Para ello, se está suponiendo que la frecuencia de las palabras tiene una incidencia directa en el tema más importante dentro del grupo, es decir que, si aparece el nombre de una empresa, se supone que la alta frecuencia del uso del nombre de la empresa en los comentarios y con base en la naturaleza de los comentarios que se descarga (comentarios relacionados a competencia desleal) se puede inferir que esa empresa pueda tener una percepción relacionada a competencia desleal.

Todos estos resultados se consolidaron en un aplicativo; se puede filtrar por grupo y por periodo de tiempo que se desea visualizar, adicionalmente se creó un gráfico de torta el cual muestra cuántos comentarios fueron analizados para el grupo o para el mes que el usuario haya decidido filtrar. El aplicativo muestra la información de cada uno del grupo como se presentados en la Figura 8. Esta última gráfica busca dar una visión de la importancia relativa del grupo de referencia que se está visualizando con respecto a la totalidad de los grupos. Lo anterior con el fin de mostrar si aparece un término dentro de la nube de palabras, la importancia pueda ser apoyada con respecto a cuantos comentarios representa ese grupo frente al total. En ella también se puede ver el gráfico de frecuencias de palabras dentro del subgrupo relativo al filtro seleccionado.

*Figura 8: Captura de pantalla de la primera vista del reporte*


*Fuente: elaboración propia*

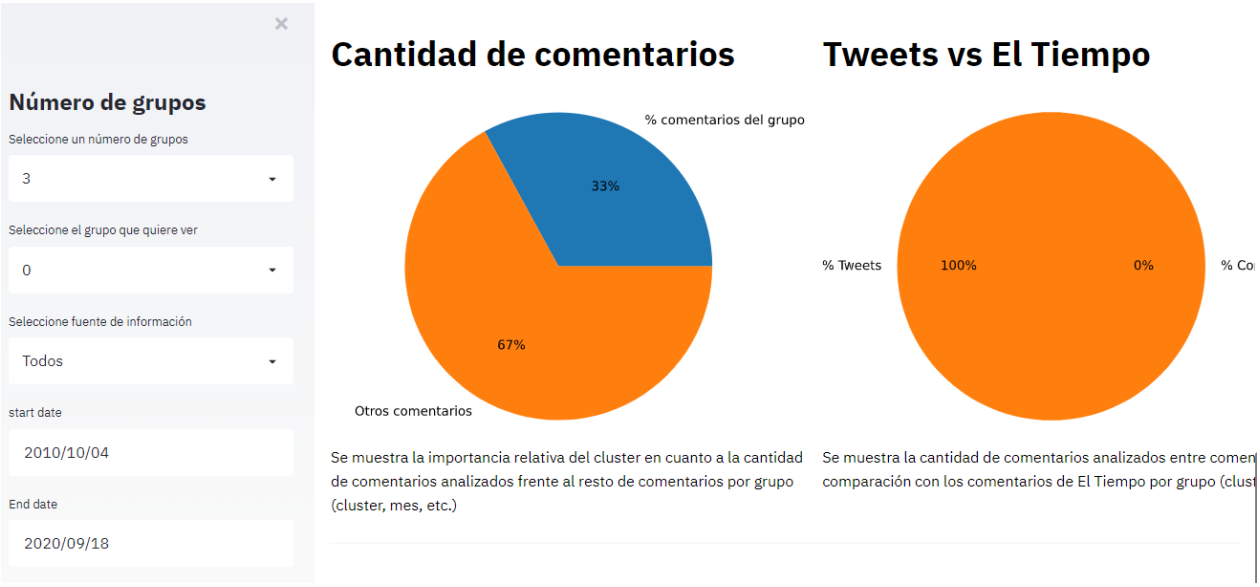
La segunda parte del aplicativo muestra el gráfico de co-ocurrencia la cual se refiere el grado de conexión que tiene cada palabra con respecto a las demás de acuerdo a su lejanía. Estas corresponden a la utilización conjunta de palabras o unidades léxicas que de acuerdo con su frecuencia se puede determinar una conexión mayor frente a otras palabras. Por ejemplo, las palabras competencia y desleal pueden ser usadas en dos contextos relativamente separados pero la unión de estas palabras cobra sentido dentro de otro contexto. De tal modo, la coocurrencia entre las palabras va a ser alta si un término sele ir acompañado de otro término. Esta también puede tener una co-ocurrencia alta si un término por una distancia de n palabras, suelen escribirse dentro de la misma oración. Un ejemplo de lo anterior es la oración “la autopista está cerrada”, las palabras “autopista” y “cerrada” no están juntas, pero dentro del sentido de la oración, ambas están conectadas. La co-ocurrencia entonces sería con una distancia de dos palabras La Figura 9 muestra la gráfica de co-ocurrencias de la unión entre las palabras dentro de un gráfico de conexiones.

*Figura 9: Captura de pantalla de la segunda vista del reporte*


*Fuente: elaboración propia*

Por su parte, la última parte muestra la cantidad de comentarios del subgrupo de referencia y la cantidad de tweets frente a la cantidad de comentarios de El Tiempo. Cabe resaltar que la cantidad de comentarios de El Tiempo es muy pequeña con respecto a la cantidad de tweets, por lo que esta información puede tener poca injerencia en los resultados y ser una su mayoría provenientes de Twitter ( Ver Figura 10)

*Figura 10 Captura de pantalla de la tercera vista del aplicativo*




*Fuente: elaboración propia*

**Conclusiones y recomendaciones**

1. La descarga de tweets se hace por medio de la librería tweepy la cual es la librería oficial de twitter para conectarse con sus servicios desde Python. Esta librería tiene las limitaciones establecidas por twitter para descarga de información, por lo que dicha información depende enteramente de las políticas de twitter.

2. Se probaron diferentes modelos de los cuales la mayoría corresponden a modelos para filtrar la información. Sin embargo, es decisión del usuario final en cuántos grupos desea dividir los comentarios y el filtro que desea visualizar.
3. El aplicativo final se encuentra desplegado en un servidor para uso interno en el DNP. Dentro del aplicativo el usuario tiene la posibilidad de realizar tantos filtros como lo considere de acuerdo a fecha, fuente (Twitter y El Tiempo) y grupo de referencia. Así mismo, puede elegir cuántos grupos desea tener para realizar el análisis.

**Socialización**

Se ha socializado con la DIDE y con el director de la DDD.