laSalle
UNIVERSITAT RAMON LLULL

TFG

Detección de *tweets* fraudulentos

Jorge Melguizo
Tutora: Elisabet Golobardes

Departamento de ingeniería informática

Curso académico 2017/18

Palabras clave

Data Science, Tweeter, Trust Tweets, Text Classification, Sentiment Analysis.

Abstract

Para el final

Resumen

La población cada vez es más exigente, en el mundo del periodismo esto se resume en inmediatez a la hora de publicar una noticia. Para ello los periodistas recurren cada vez más a las redes sociales como fuente de información y con la exigencia de ser los primeros en publicar los sucesos, no contrastan la información leída. Si lo combinamos con que gran parte de la información que está en las redes sociales suele ser falsa hace que un periodista publique una noticia falsa y provoco que la sociedad se crea una mentira por el hecho de haberlo visto en televisión o leído en algún medio de información. Este proyecto pretende dar caza a mentiras en las redes sociales pero acotándolo al mundo de los deportes.

Extender

Índice

1. Introducción

1.1. Contexto

En la era de la tecnología en la que vivimos cada vez se da más importancia a las redes sociales. Ya no se usan las redes sociales exclusivamente para hablar y compartir cosas con amigos, ahora también lo usan las marcas para hacer campañas publicitarias, los informativos para adelantar noticias y los equipos para informar a sus seguidores. Además de la importancia que de por sí tienen, actualmente, la sociedad informatizada en la que nos hemos adentrado en el último siglo, ha hecho que el interés por el análisis de la opinión y veracidad de la información crezca exponencialmente en estos últimos años.

Los seres humanos siempre han querido conocer de inmediato los acontecimientos que suceden a su alrededor. Para ello antiguamente se recurría de los diarios pero solo te decían lo que había sucedido el día anterior, después fue el tiempo de la televisión con sus informativos dos veces al día, a continuación apareció internet y los diarios y cadenas de televisión se aprovecharon de ello para ganarle la partida al otro e informar de los hechos en cuestión de minutos.

A partir de esta guerra interna que hay en el mundo del periodismo de informar el primero y tener imágenes de todo el mundo, actualmente, los periodistas, se ayudan de las redes sociales para enterarse de lo que sucede en el mundo. Esto es una gran idea pero, en el mundo de internet, siempre hay gente mintiendo.

Dentro de este cambio que han sufrido las redes sociales cada vez es más importante saber en quien se puede confiar y que *posts* hay que ignorar. Quien sube estás noticias son los *trolls*, gente que publica falsos *posts* simplemente para divertirse o porque no saben sobre lo que estan escribiendo o *retweeteando*.

De los nombrados *trolls* hay de muchos tipos. Una de las clasificaciones es:

1. **El principiante**. Es aquel que se abre un perfil, amparado en el anonimato, y con un número de seguidores insignificante. Generalmente, su trolleo pasará inadvertido y no alcanzará su propósito ni de lejos.
2. **El estratega**. Tiene claro el objetivo, contactará con usuarios con identidad digital, buen nivel de penetración, credibilidad y un buen número de seguidores. Hará

que ellos se encarguen de propagar su crítica y sin duda pueden conseguir hacer ruido en las redes.

3. **El sarcástico**. Te puede sacar de quicio en un momento dado. Le encanta liarte. Su finalidad suele ser la de provocar para que no te olvides de él y por supuesto que no te lo calles. Necesita que cuentes qué hace y que señales quién es a los demás.

4. **El sádico**. Su objetivo es disfrutar con tu dolor y humillación y no parará hasta saciar su sed con tu sufrimiento. Es capaz de sacar de contexto cualquier acción que hagas por el mero hecho de ver cómo los demás hacen carnaza con ello. Lo que peor que puedes hacer es contestar a su provocación pues esa será la señal de su ataque.

5. **El arrepentido**. Es el que finalmente saca a pasear su conciencia y le sobreponen sus acciones. Pero no te fíes, casi siempre vuelve.

6. **El cansino**. Aparece sorpresivamente de la nada metiéndose en una conversación y no para de darte palique hasta límites insospechados. Debes pararle los pies o se convertirá en tu peor pesadilla. Bloquearlo puede ayudar, aunque hay gestores de redes que permiten saltarse el bloqueo.

7. **El omnipresente**. No concibe otra forma de estar en redes sociales que no sea trolleando a diestro y siniestro, especialmente a usuarios con muchos seguidores. Su finalidad, en el caso de que no tenga credibilidad, es la del egocentrismo.

8. **El lerdo**. Va de listo utilizando un usuario falso y publica el mismo mensaje en su cuenta personal. Le pillas y le hundes, a la par que regalas un buen momento a tus seguidores.

9. **El frustrado**. Conocido también como hater. La difamación, el insulto son su sistema de trolleo. Un desencadenante fundamental para ser un hater es el odio. Su frustración y decepción, en algunos casos interna, le hacen imponer su criterio sobre cualquier otro. Es más, harán lo imposible por demostrar que quien no piensa como él es despreciable y humillable. Cualquier cosa que hagas estará mal y será reprochable. Aunque tengas paciencia muy probablemente acabarás buscando un abogado.

10. **El oportunista**. Aprovecha el trending topic del momento para insultar a quienes estén relacionados con el tema o para hablar de sus temas recurrentes incluyendo el hashtag sin que éste tenga relación alguna con ellos. Todo por intentar tener un poco de visibilidad.

11. **El suplantador**. Simula ser otro usuario para intentar dañar su imagen con mensajes que le perjudican o, si se trata de alguien popular, con el fin de captar muchos seguidores para terminar vendiendo la cuenta.

12. **El paródico**. No intenta suplantar a otro usuario, sino parodiarle, unas veces con el fin de burlarse de él de forma más o menos sana y otras con la intención de difamarle.

13. **El zombi**. Es una cuenta creada o comprada por un troll con el fin de lanzar mensajes automatizados para atacar a alguien.

14. **El vampiro**. Es uno de los más peligrosos. Se alimenta del sufrimiento de sus víctimas y en muchos casos vive obsesionado con ellas, monitorizando absolutamente todas sus actividades dentro y fuera de las redes con la finalidad de lanzar sus ataques. Es un delincuente y cada día hay más casos en los tribunales.

15. **El cazador de trolls**. Hay usuarios que terminan vengándose y trollean a sus propios trolls, ya sea desde sus cuentas personales o creando perfiles específicos para la ocasión.

1.2. Motivaciones

Las opiniones y sentimientos de las personas siempre han sido una valiosa fuente de información. Por ejemplo, las grandes empresas como Apple o Samsung necesitan conocer el impacto público que producen y mantener controlada al día la opinión pública que será su medidor de éxito de su venta ya que no les vale con vender mucho una vez si no que necesitan vender mucho y muchas veces. Esto les hace recolectar toda la información posible de los clientes ya sea vía mail, redes sociales o cualquier otra forma de saber la opinión de la gente sobre de sus productos. Esto también se aplica a políticos para saber si están enfocando de manera correcta su campaña electoral e incluso las empresas de marketing siempre han tomado la opinión social como el centro de sus actividades. Incluso se ha utilizado esta información como estudios de mercado para venderlos a las empresas y no son más que una recolección de datos.

El problema de estos textos que depositamos en nuestras redes sociales es la fiabilidad de la fuente. Cualquier persona que tenga una cuenta puede acceder y comentar sobre el tema que el quiera y ahí es donde aparecen los trolls comentados anteriormente. A veces son comentarios inofensivos que no molestan a nadie como serian comentarios no verificados del estilo de 'Yo cocino mejor que Ferran Adria' donde no hay ninguna maldad sino que se podría considerar como un comentario irónico o gracioso.

El motivo principal para la elección de este proyecto ha sido raíz a las mentiras que se llevaron a cavo durante el atentado de Barcelona en las ramblas yo era una de las personas que estaban de vacaciones pero que tenía familia en Barcelona y que la gente con o sin maldad se dedicaba a retwitear todo sin saber ni siquiera si era verdad o se trataba de una fuente fiable y lo único que hacían era provocar pánico, angustia y desconcierto a miles de familias.

1.3. Objetivos Trabajo Final de Grado

Con el auge de las redes sociales en la última década, los usuarios de redes sociales tales como Twitter, Facebook o Instagram no paran de crecer, y en ellos los usuarios no paran de volcar y compartir opiniones y sentimientos sobre cualquier tema, evidentemente lo que más se comenta es lo que suceda en la actualidad de aquel momento, creando así cantidades inmanejables de datos. El problema de estos datos es el analizarlos y guardarlos de forma estructurada, que sea útil y que nos permita analizar los resultados de forma que éstos nos lleven a conclusiones beneficiosas.

Este proyecto se centrará en cazar a los *trolls* de *Tweeter* que intentan engañar en el mundo de los deportes. Se pretende clasificar *tweets* falsos y verdaderos clasificando estos primero en el deporte al que se vincula el *tweet* (fútbol, baloncesto, balonmano, golf, rugby, etc), según el tono en el que está escrito con el uso de análisis del sentimiento, esto pretende no solo decir si el texto es positivo o negativo sino una clasificación más refinada sobre tipo de sentimiento con el que esta escrito. Finalmente, con estas técnicas en el tweet analizado y tanto anteriores del usuario como tweets con sus hashtags y, habiendo aplicado también una formula propia de fiabilidad, se querrá dar un porcentaje de fiabilidad del tweet.

Como ya se ha comentado nos centraremos en la sección de noticias donde hay más mentiras (los deportes) pero, se pretende hacer de tal manera que, solo cambiando los diccionarios de palabras usadas, se puedan aplicar las mismas técnicas y nos permita conocer .ª tiempo real"la fiabilidad de los tweets que se leen o que se están volcando en twitter en ese momento.

1.4. Estructuración de la memoria

Para el final

2. Trabajo Relacionado

Este trabajo se trata de un proyecto muy ambicioso y como tal abarca un gran abanico de temáticas, por lo tanto, hay mucho trabajo relacionado con partes del proyecto. La gran diferencia es que este proyecto esta en español (la mayoría están en inglés) y que no se centra solo en un concepto.

Empezamos con un proyecto que se essta llevando a cabo desde el 2014 hay un proyecto financiado por la UE llamado ***Pheme***. El cual es llamado el detector de mentiras de internet. Pheme dice pretender solucionar el cuarto problema del Big Data; la veracidad. Analizará las publicaciones sociales en redes como Twitter y Facebook, las comparará y las contrastará con otras fuentes de información, con el fin de establecer su veracidad y, en el caso de que se clasifique como bulo, señalar la razón (estableciendo si estamos ante una simple especulación, una controversia, un ejercicio deliberado de desinformación, etc).

Luego hay muchos trabajos relacionados respecto a la psicología que hay que aplicar para detectar perfiles falsos en twitter como un articulo que presento el diario abc [1] y también sus características. [2]. Incluso hay quien le dio la vuelta a la tortilla y decidio buscar las características de los tweets confiables [3].

De trabajos teóricos para detectar noticias falsas hay muchos [4] [5] [6] pero no los hay de aplicarlo que se haga automáticamente en una maquina con un programa, el factor humano es clave para ellos. También respecto a los tipos de troll que existen [7] [8] incluso papers donde se hablan de las llamadas *fake news* [9].

Después de todos estos trabajos teóricos también podemos encontrar algunas implementaciones para encontrar bots [10] e incluso herramientas para encontrar seguidores

[1] http://www.abc.es/tecnologia/redes/20140320/abci-faketwitter-saber-twitter-falsos-201403191326.html
[2] http://www.outono.net/elentir/2013/02/02/twitter-5-formas-de-identificar-a-un-troll
[3] http://www.josemorenojimenez.com/2012/03/20/caracteristicas-de-un-tweet-confiable/
[4] https://www.entrepreneur.com/article/292342
[5] http://www.eltiempo.com/tecnosfera/novedades-tecnologia/como-identificar-noticias-falsas-en-redes-sociales
[6] http://www.abc.es/tecnologia/redes/20131210/abci-como-detectar-noticia-falsa-201312092125.html
[7] https://blogs.elconfidencial.com/tecnologia/elclubdelalucha/2015-05-09/troll-redes-sociales-ciberacoso_790558/
[8] http://www.elmundo.es/f5/comparte/2017/10/19/59e77fbd22601db82d8b459f.html
[9] http://www.kdd.org/exploration_files/19-1-Article2.pdf
[10] https://botometer.iuni.iu.edu/#!/

falsos [11].

Este tema de encontrar trolls y de la preocupación que hay de que la tecnologia no acabe con las personas es una area de interes donde se han introducido grandes empresas como Google con su *Troll detection* [12], la librería de python de referencia Scickit-Learn [13] e incluso el gran medio de comunicación internacional BBC habla al respecto [14].

Por otra parte, esta clasificación de los usuarios en trolls, tambien hay quien,contando que hay trolls agresivos, y no que solo dicen mentiras si no que intentan incordiar a los usuarios o que aplican lo que ahora se conoce como Cyber Bulling, dado esto, se han escrito articulos para la detección de estos usuarios [15].

[11] https://www.genbeta.com/redes-sociales-y-comunidades/7-herramientas-para-detectar-si-tienes-seguidores-falsos
[12] https://www.theverge.com/2017/2/23/14713496/google-jigsaw-perspective-software-ai-machine-learning-developers
[13] http://blog.kaggle.com/2012/09/26/impermium-andreas-blog/
[14] http://www.bbc.com/news/technology-38181158
[15] https://academic.oup.com/jigpal/article-abstract/24/1/42/2893010?redirectedFrom=fulltext

3. Sports Fake Detector

3.1. Tecnologías utilizadas

3.1.1. MongoDB

MongoDB es un sistema de bases de datps NoSql de código abierto orientado a documentos. En lugar de almacenar la información en tablas como se haría en una base de datos relacional, MongoDB guarda la información en forma de documentos JSON, haciendo que la integración de este sistema de información sea mucho más simple y rápido que otros sistemas de bases de datos.
MongoDB ha sido muy útil para almacenar de manera eficiente grandes conjuntos de tuits. Para trabajar con este sistema de bases de datos se ha empleado la líbreria PyMongo, que es la distribución para Python recomendada en la web oficial de MongoDB y que contiene todas las herramientas necesarias. Para el proyecto se ha empleado la versión 3.0.4 de MongoDB.
MongoDB es un sistema fácil de utilizar y que funciona muy bien para guardar textos pero vamos a ver una tabla donde se puede comparar una base de datos como MongoDB con una SQL[16].

Feature	**NoSQL Database**	**SQL Database**
Performance	High ✓	Low
Reliability	Poor	Good ✓
Availability	Good	Good
Consistency	Poor	Good ✓
Dara storage	Optimized for huge data ✓	Medium sized to large
Scalability	High ✓	High but expensive

[16]Obtenida en `https://blog.pandorafms.org/es/bases-de-datos-nosql/`

3.1.2. Pymongo 3.7.1

Pymongo es una librería de Python para poder conectarnos a una base de datos MongoDB. Para utilizar esta librería se utilizará una clase propia.

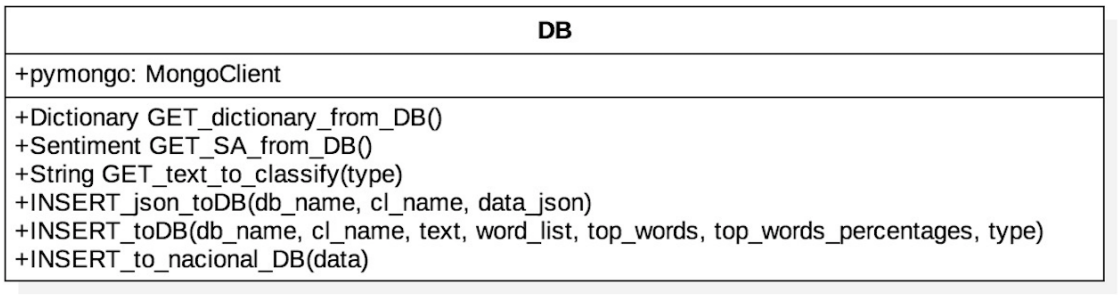

Figura 1: Clase DB utilizada para la comunicación entre el programa y la base de datos en MongoDB

3.1.3. AVL Tree

Todas las palabras están guardadas en una base de datos MongoDB pero con la finalidad de hacer la búsqueda más rápida una vez se consultan por primera vez se quedan guardadas en un árbol propio, concretamente en un AVL tree.

Un árbol es un tipo abstracto de datos (TAD) muy utilizado que imita la estructura jerárquica de un árbol, con un valor en la raíz y subárboles con un nodo padre, representado como un conjunto de nodos enlazados.

La ventaja que tiene este tipo de árboles (AVL) es que están siempre equilibrados de tal modo que para todos los nodos, la altura de la rama izquierda no difiere en más de una unidad de la altura de la rama derecha o viceversa. Gracias a esta forma de equilibrio (o balanceo), la complejidad de una búsqueda en uno de estos árboles se mantiene siempre en orden de complejidad O(log n). El factor de equilibrio puede ser almacenado directamente en cada nodo o ser computado a partir de las alturas de los subárboles.

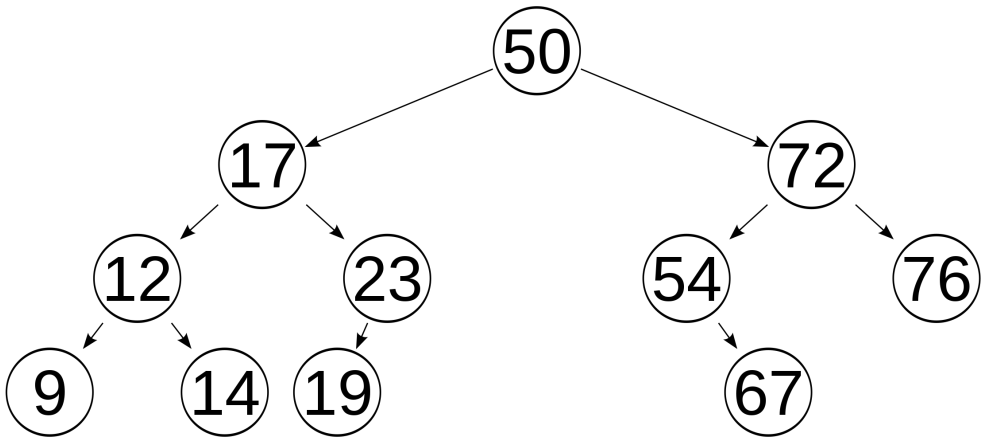

Figura 2: Ejemplo AVL Tree

Como se puede observar en la figura anterior, un árbol es unicamente una estructura de datos ordenada. En la figura es un orden numérico (de menos a más valor) y el árbol utilizado en este trabajo está ordenado alfabéticamente. Con este árbol tendremos un coste inicial de creación pero luego el coste medio de la búsqueda de palabras será mucho menor.

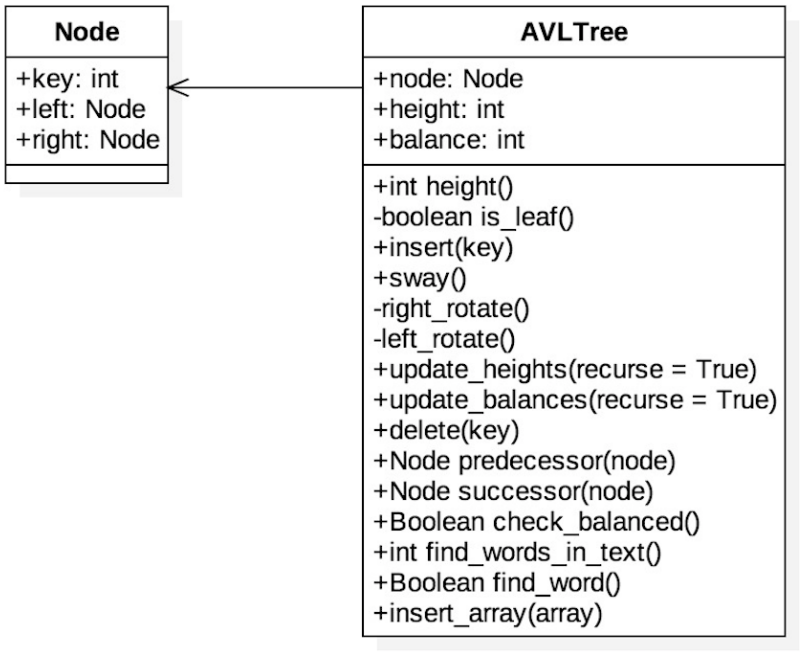

Figura 3: Diagrama de la clase propia AVL-TREE utilizada para la interacción con el árbol

3.2. Fase 0 - Extracción de la información y creación de los datasets

3.2.1. Concepto

Esta primera fase o fase inicial se irá modificando al largo de la implementación según se necesite. Pretende englobar tanto la extracción de los tweets como la creación de los diccionarios.

Se pretende así que cada deporte tenga tres diccionarios asociados (palabras excluyentes, palabras vinculantes principales y palabras vinculantes secundarias). También se tendrán diccionarios para el análisis de las frases; palabras vacías, que son aquellas palabras que no aportan nada a la hora de clasificar el texto como podrían ser artículos y preposiciones.Por último se tendrá un diccionario asociado a cada uno de los sentimientos a analizar; alegría, amor, enfado, miedo, sorpresa y tristeza.

3.2.2. Software

El esquema de esta fase se puede ver en la siguiente imagen:

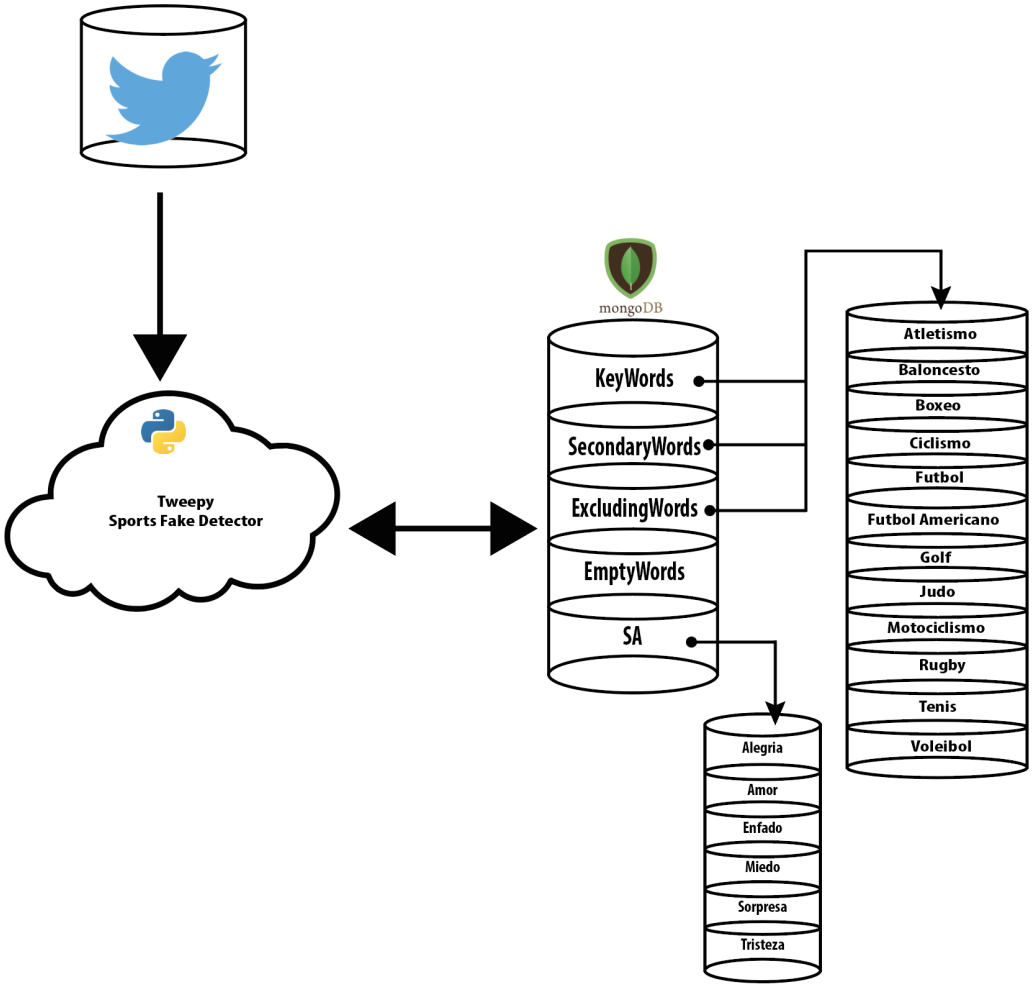

Figura 4: Estructura de extracción de tweets y interacción con diccionario

Donde se puede ver que se extrae la información de twitter a través de la librería tweepy, se proceso con el detector de mentiras y se interacciona con la base de datos donde previamente se habían guardado los diccionarios.
Las tablas de la base de datos son las siguientes:

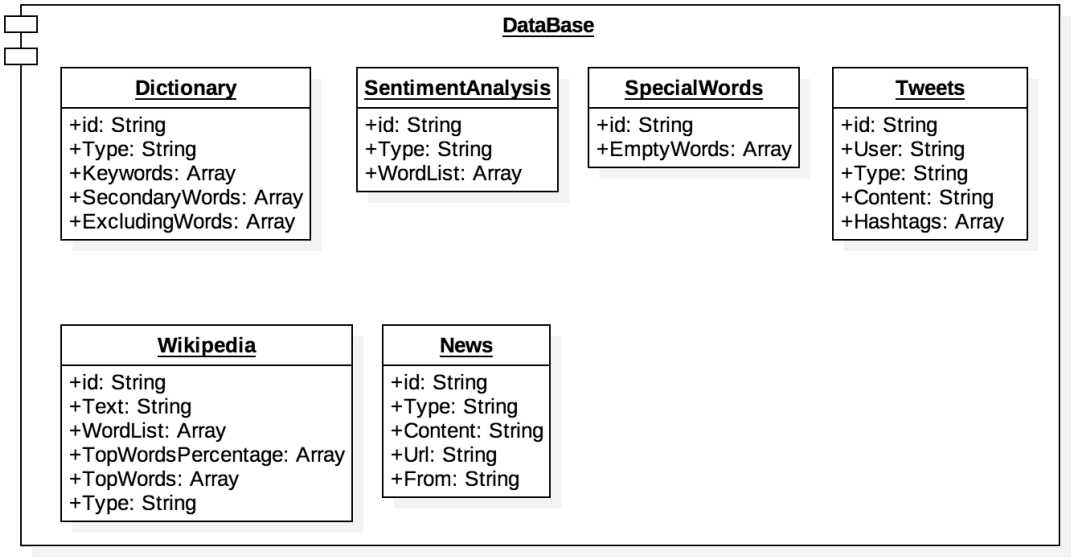

Figura 5: Tablas en la base de datos del proyecto

3.2.2.1 Tweepy 3.5

Tweepy es una librería de código abierto para Python que incluye todo el conjunto de funciones necesarias para comunicar con Twitter mediante las API's definidas por este. Las funciones definidas por Tweepy simplifican la conexión y búsquedas con Twitter. Por ejemplo, toda la conexión con Twitter debe estar certificada con un autor y, mientras que por defecto habría que configurar esta conexión mediante otra librería como sería Python-Auth y establecer cada conexión manualmente, Tweepy simplifica esto con unas funciones que simplemente esperan está autenticación como parámetro para poder configurarlo automáticamente. Estos parámetros son 4 tokens necesarios y eb el caso de la búsqueda solo tenemos que indicarle los parámetros que solicita Twitter, toda la complejidad de las conexiones la trata internamente simplificando el trabajo inmensa,ente. Para el proyecto se ha empleado la versión 3.3 de Tweepy.

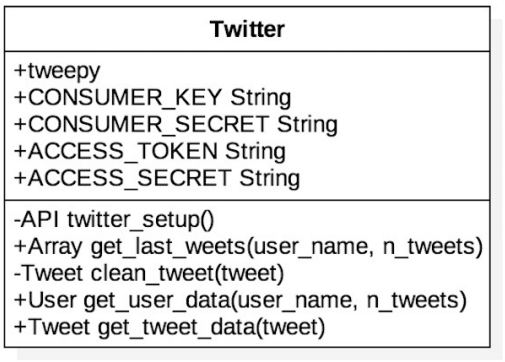

Figura 6: Diagrama de la clase propia Twitter encargada de la comunicación con twitter mediante la librería tweepy

3.2.2.2 Wikipedia 1.4.0

Está librería se empezó a usar con la finalidad de crear diccionarios automáticamente, simplemente coger todas las palabras que salen de una búsqueda en wikipedia, excluir todos los artículos, proposicionales y palabras que no aportan valor de clasificación.

De todas las palabras obtenidas se cogían el 30 % de las palabras más repetidas. Esto permitió empezar a hacer pruebas de los algoritmos pero se descarto por falta de precisión y se empezó a hacer un diccionario propio que ganó riqueza por algunas palabras encontradas gracias a este algoritmo.

3.3. Fase 1 - Clasificación de texto según temática

3.3.1. Concepto

TODO

3.3.2. Software

TODO: Hablar de las librerías python utilizadas, diagrama ...

3.3.2.1 Stemmer

Por razones gramaticales, los documentos van a utilizar diferentes formas de una palabra, como jugar, juegan y jugaron o jugador y jugadores. Además, hay familias de palabras derivadas relacionadas con significados similares, como vivir y convivir. En muchas situaciones, parece que sería útil buscar una de estas palabras para devolver documentos que contengan otra palabra en el conjunto.

El objetivo tanto de la derivación (stemmer) como de la lematización (lemmatization) es reducir las formas flexivas y, a veces, las formas derivadas de una palabra a una forma básica común.

Sin embargo, las dos palabras difieren en su forma de hacerlo. Stemming usualmente se refiere a un crudo proceso heurístico que elimina los extremos de las palabras con la esperanza de lograr este objetivo correctamente con un alto porcentaje, y a menudo incluye la eliminación de los afijos derivacionales. La lematización generalmente se refiere a hacer las cosas correctamente con el uso de un vocabulario y un análisis morfológico de las palabras, normalmente con el objetivo de eliminar solamente las terminaciones y devolver la forma base o diccionario de una palabra, lo que se conoce como el lema (lemma).

Como este trabajo esta en castellano, una lengua muy rica en su vocabulario y derivaciones, se ha optado por usar un Stemmer y no un lematizador dado que se obtienen mejores resultados.

```
from nltk.stem import SnowballStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
def stem(word):
    stemmer = SnowballStemmer('spanish')
    return stemmer.stem(word)
new_text = "Jugando con mi sobrino nos dimos cuenta de lo divertido que es el escondite"
text_tokens = word_tokenize(new_text)
stemmed = []
for item in text_tokens:
    stemmed.append(stem(item))
print()
print('Input: ', new_text )
print('Output:', stemmed )
```

Figura 7: Código ejemplo utilización VADER

```
Input:  Jugando con mi sobrino nos dimos cuenta de lo divertido que es el escondite
Output: ['jug', 'con', 'mi', 'sobrin', 'nos', 'dim', 'cuent', 'de', 'lo', 'divert', 'que', 'es', 'el', 'escondit']
```

Figura 8: Output ejemplo utilización VADER

3.3.2.2 Naive Bayes

El clasificador Naive Bayes es un clasificador probabilístico que se basa en el teorema de Bayes con suposiciones de independencia. Es una de las técnicas de clasificación de texto más básicas con varias aplicaciones en la detección de spam en el correo electrónico, clasificación de correo electrónico personal, categorización de documentos, detección de lenguaje y detección de sentimiento.

Como se ha comentado se basa en el teorema de Bayes así que primero lo comentaremos y se usará como ejemplo la clasificación de texto. Se usa para predecir la probabilidad condicional de que un documento pertenezca a una clase $P(c_i|d_j)$ a partir de la probabilidad de los documentos dada la clase $P(d_j|c_i)$ y la probabilidad a priori de la clase en el conjunto de entrenamiento $P(c_i)$

$$P(c_i|d_j) = \frac{P(c_i)P(d_j|c_i)}{P(d_j)}$$

Figura 9: Formula Naive Bayes

Dado que la probabilidad de cada documento $P(d_j)$ no aporta información para la clasificación, el término suele omitirse. La probabilidad de un documento dada la clase suele asumirse como la probabilidad conjunta de los términos que aparecen en dichos documentos dada la clase y se calculan como:

$$P(d_j|c_i) = \Pi_{t=1}^{|V|} P(w_t|c_i)$$

Figura 10: Formula Multinomial Bayes

3.3.2.3 Diccionario

TODO

3.4. Fase 2 - Análisis del sentimiento

3.4.1. Concepto

TODO

3.4.2. Software

TODO: Hablar de las librerías python utilizadas, diagrama ...

3.4.2.1 Vader Sentiment 3.3

VADER (Valence Aware Dictionary and sEntiment Reasoner) es un léxico y una biblioteca de análisis de sentimiento basada en reglas. Sirve como herramienta de análisis de sentimientos basada en reglas que está específicamente adaptada a los sentimientos expresados en las redes sociales.

Es completamente de código abierto bajo la [Licencia MIT]. VADER puede incluir el sentimiento de los emoticonos (por ejemplo, :-)), los acrónimos relacionados con el sentimiento (por ejemplo, LOL) y la jerga (por ejemplo, meh).

El método vaderSentiment() devuelve los valores para representar la cantidad de sentimiento negativo, positivo y neutral y también calcula el valor de sentimiento compuesto como un valor con signo para indicar la polaridad del sentimiento general.

A continuación se muestra un ejemplo de utilización de VADER:

```
sentences = ["Amo a Leo Messi es el mejor del mundo <3",  # positive sentence example
             "Como es posible que me guste tanto Messi? Amor etertno!",  # punctuation emphasis handled correctly (sentiment intensity adjusted)
             "Cristiano esta sobrevalorado, no es buen jugador y nunca lo fue",  # negation sentence example
             "El golf no esta mal, meh.",  # positive sentence
             "El golf es un deporte muy aburrido que no me gusta nada.",  # negated negative sentence
             "No se si me gusta menos Cristiano Ronaldo o es peor el golf", # mixed negation sentence
             "Hoy jugaron fatal, no entiendo que apso lol", # mixed sentiment example with slang and constrastive conjunction "but"
             "Make sure you :) or :D today!",  # emoticons handled
             "Si viese a mi amor Leo Messi 💘 solo 💋 no hay palabras😁",  # emojis handled
             "Los lakers no estan mal del todo" , # Capitalized negation
             "Hola"  # Neutral
             ]
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
for sentence in sentences:
    vs = analyzer.polarity_scores(sentence)
    print("{:-<70} {}".format(sentence, str(vs)))
```

Figura 11: Código ejemplo utilización VADER

```
Amo a Leo Messi es el mejor del mundo <3---------------------------- {'neg': 0.0, 'neu': 0.734, 'pos': 0.266, 'compound': 0.4404}
Como es posible que me guste tanto Messi? Amor etertno!------------- {'neg': 0.0, 'neu': 0.677, 'pos': 0.323, 'compound': 0.6476}
Cristiano esta sobrevalorado, no es buen jugador y nunca lo fue----- {'neg': 0.196, 'neu': 0.804, 'pos': 0.0, 'compound': -0.296}
El golf no esta mal, meh.------------------------------------------- {'neg': 0.467, 'neu': 0.533, 'pos': 0.0, 'compound': -0.3612}
El golf es un deporte muy aburrido que no me gusta nada.------------ {'neg': 0.167, 'neu': 0.833, 'pos': 0.0, 'compound': -0.296}
No se si me gusta menos Cristiano Ronaldo o es peor el golf--------- {'neg': 0.167, 'neu': 0.833, 'pos': 0.0, 'compound': -0.296}
Hoy jugaron fatal, no entiendo que apso lol------------------------- {'neg': 0.422, 'neu': 0.37, 'pos': 0.207, 'compound': -0.4404}
Make sure you :) or :D today!--------------------------------------- {'neg': 0.0, 'neu': 0.294, 'pos': 0.706, 'compound': 0.8633}
Si viese a mi amor Leo Messi 💘 solo 💋 no hay palabras😁------------- {'neg': 0.105, 'neu': 0.571, 'pos': 0.324, 'compound': 0.6808}
Los lakers no estan mal del todo------------------------------------ {'neg': 0.268, 'neu': 0.732, 'pos': 0.0, 'compound': -0.296}
Hola---------------------------------------------------------------- {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

Figura 12: Output ejemplo utilización VADER

3.4.2.2 TextBlob 0.15.1

TextBlob es una librería de Python (2 y 3) para procesar datos textuales. Proporciona una API simple para adentrarse en tareas comunes de procesamiento del lenguaje natural (NLP), como etiquetado parcial, extracción de frase nominal, análisis de sentimiento, clasificación y traducción.

En esta fase se ha utilizado esta librería para traducir tweets de otros idiomas y para la clasificación del sentimiento se escogió esta librería porque que ofrece una API simple para acceder a sus métodos y realizar tareas NLP básicas.

3.4.2.3 NLTK 3.2.2

NLTK (Natural Language Toolkit) es una plataforma líder para construir programas de Python para trabajar con datos de texto natural. Proporciona interfaces fáciles de usar a más de 50 recursos léxicos como WordNet, junto con un conjunto de librerías de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico, envoltorios para bibliotecas de PNL de fuerza industrial, y un foro de discusión activo.

Contiene una completa documentación (API). Está disponible para Windows, Mac OS X y Linux. NLTK es un proyecto gratuito y de código abierto que ha sido llamado ünа herramienta maravillosa para enseñar y trabajar en lingüística computacional usando Pythonz ünа biblioteca increíble para jugar con el lenguaje natural".

3.4.2.4 Diccionario

TODO: SENTIMENT ANALYSIS WORDS

3.5. Fase 3 -

3.5.1. Concepto

3.5.2. Software

4. Experimentación

4.1. Problemática lenguaje natural

El procesamiento del lenguaje natural (PLN o NLP en sus siglas anglosajonas) es una rama de la Inteligencia Artificial encargada de estudiar métodos de comunicación entre máquina y hombre a través del lenguaje natural, es decir, el lenguaje empleado de forma habitual en una conversación escrita u oral entre personas.

El lenguaje natural presenta muchas características que lo hacen un verdadero reto para las ciencias de la computación. Como suele pasar en el mundo de la Inteligencia Artificial, una tarea que puede parecer para una máquina. Aspectos como la ambigüedad, la espontaneidad, la falta de fluidez, las referencias y las abreviaturas son difíciles de procesar.

4.1.1. Homonimia

La homonimia es la cualidad de dos palabras, de distinto origen y significado por evolución histórica, que tienen la misma forma, es decir, la misma pronunciación o la misma escritura.

Un ejemplo estaría en la palabra **vela**:

1. Acción de velar; cilindro de cera con una mecha para iluminar. (Ambos sentidos relacionados con el verbo velar.
2. Tela grande que aprovecha la fuerza del viento, especialmente en un barco.

En castellano de homonimias tenemos diferentes clases:

1. **Homónimos lexicales**: los que pertenecen a la misma categoría gramatical: onda y honda, botar y votar, haya y aya, ojear y hojear.
2. **Homónimos gramaticales**: los que no pertenecen a la misma categoría gramatical: cabe verbo y cabe preposición, o los que perteneciendo a la misma categoría gramatical se diferencian en alguna marca morfemática: el pez, la pez; el orden, la orden.
3. **Homónimos léxico-gramaticales**: los que se han formado a través de un cambio de funciones: poder (verbo) poder (sustantivo)

4. **Homónimos morfológicos**: cuando se producen diferentes formas de una sola palabra: decía primera y tercera personas del pretérito imperfecto de indicativo; o se dan formas correspondientes de palabras diferentes: fui (de ser e ir); ve (de ir y de ver), etc.

4.1.2. Polisemia

La polisemia, en lingüística, se presenta cuando una misma palabra o signo lingüístico tiene varias acepciones o significados. Una palabra polisémica es aquella que tiene dos o más significados que se relacionan entre sí.

Cabe resaltar que la polisemia puede surgir por diversos motivos. Por un lado, el vocabulario figurado produce polisemia por medio de las metáforas y las metonimias. Por ejemplo: los brazos de un río, las patas de una mesa. La especialización y el lenguaje técnico también atribuyen un significado específico a ciertos términos (como en el caso del ratón en la informática).
La influencia extranjera y las modificaciones de aplicación son otras condiciones que favorecen la polisemia: una muestra de esto es el vocablo botón que nació con la indumentaria y luego pasó a utilizarse también en los artefactos electrónicos.
Un ejemplo estaría en la palabra **cabo**:

1. (masculino) Punta de tierra que penetra en el mar.
2. (masculino/femenino) Escalafón militar.
3. (masculino) Cuerda en jerga náutica.

4.1.3. Anáfora y elipses

La anáfora se puede definir como la palabra o palabras que asumen el significado de una parte del discurso (texto) que ya se ha mencionado antes.
Un ejemplo de anáfora sería:

- Estoy de paso por Madrid. (Madrid) Es maravillosa.

Cuando el elemento sustituido aparece después del sustituto, hablamos de catáfora.
Un ejemplo de catáfora sería:

- Quería estar con Jesús; pero no lo he visto en toda la tarde.

Por otro lado tenemos las elipses que es la omisión o supresión de una o varias palabras que ya se habían mencionado antes o que se pueden presuponer o sobreentender. Esta construcción evita el uso de repeticiones innecesarias.
Un ejemplo de elipse sería:

- El concierto fue genial, la comida (fue) regular.

4.1.4. Sintaxis no normalizada

Otro problema que nos plantea el análisis de lenguaje natural es el hecho de que su estructura no está normalizada, es decir, a la hora de utilizar el lenguaje en español, no se utiliza una sintaxis concreta y bien definida, si no que a la hora de expresar una misma idea, ésta se puede estructurar de diversas maneras.

- Me encantan las mañanas de domingo.
- Las mañanas de domingo me encantan.

Ambas oraciones expresan la misma idea utilizando un orden diferente en las distintas palabras que las componen.

4.1.5. Contexto

El contexto es el conjunto de circunstancias (materiales o abstractas) que se producen alrededor de un hecho, o evento dado, que hacen que una palabra pueda tomar varios sentidos según el momento en el cual se utiliza.

Un claro ejemplo de la importancia del contexto es la ironía, a un suceso malo se puede contestar con otra frase 'mala' como sería el caso de; *¡Otra vez no!*, con una frase buena del estilo; *¡Qué bien!* donde estaríamos siendo irónicos o con una frase como; *¿Porqué a mí?*.

4.1.6. Otras circunstancias

Aparte de toda esta serie de fenómenos que se dan en la lengua formal, hay que tener que cuenta otros aspectos que se dan en ambientes más coloquiales como son las redes sociales.

En la siguiente tabla se muestran algunos ejemplos de cómo un mismo mensaje puede tener formas diferentes según el usuario y el medio en el que lo escriba.

Texto de ejemplo	**Explicación**
Messi es el mejor jugador de la historia del fútbol.	Frase correct
Mesi es el mejor jugador de la istoria del futbol;	Faltas de ortografía
MesSi es el MeJor juGAdor de la hIStoriA del fúTBol :)	Incluir mayúsculas y símbolos innecesarios
Messiiii es el mejor jugador de la historia del futbooooool	Añadir repetición de letras para darle mayor énfasis a la palabra
Mssi s l mjor jugadr d la histria dl futbol	Lenguaje SMS.
¡Vamos equipo!	Brevedad y ambigüedad del texto

4.2. Fase 0 - Extracción de la información

4.3. Fase 1 - Clasificación de texto según temática

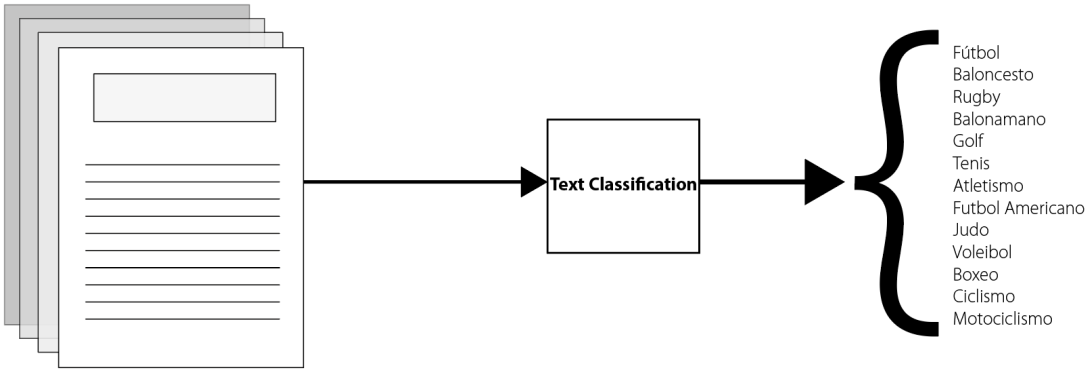

Figura 13: Idea principal del clasificador de texto

Para está primera fase primero se quiso evaluar según el contenido de Wikipedia. Se extrajeron las palabras más repetidas del contenido de cada deporte en Wikipedia. Para ello se quitaron los signos de puntuación y palabras vacías, entendiendo como palabras vacías todos los artículos, preposiciones y números. TODO: adjuntar foto resultado

Con esta técnica se pudo ver que funcionaba bien para clasificar textos grandes como artículos o noticias pero en frases cortas como son los *tweets* no se conseguía una *accuracy* que se pudiera considerar aceptable. Para mejorarlo se decidió hacer una ontología de cada uno de los deportes. En concreto se decidió clasificar las palabras que envolvían un deporte según tres criterios; *key words* que son las palabras que solo se utilizan en un deporte, sus jugadores más destacados y los nombres de sus campeonatos (ej: fútbol; Lionel Messi, bota de oro, etc), *secondary words* que son palabras que se pueden usar en ese deporte pero que pueden referir a algún otro deporte (ej: fútbol; balón, pelota, botas, etc) y por último *excluding words* que son todas las *key words* de los otros deportes y *secondary words* de otros deportes que excluyen (ej: fútbol; manillar, palo de hierro, etc).

Se intento hacer stemmer pero empeorava los resultados, lo probaremos otra vez en analisis del sentimiento se hizo con wiki, por diccionario propio y para mejorar los tiempos con diccionario propio y arbol

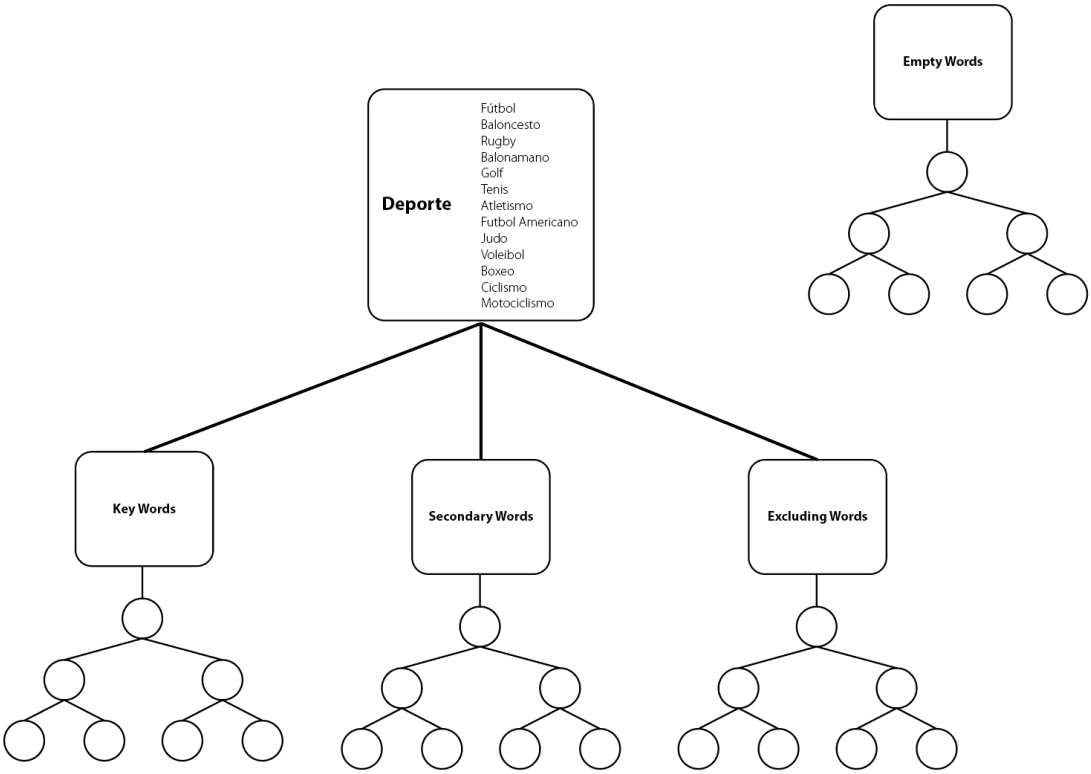

Figura 14: Esquema estructuración de los datos

TODO: TABLA COMPARATIVA DE LOS METODOS

4.4. Fase 2 - Análisis del sentimiento

El análisis de sentimiento de textos en las redes sociales (que adopta diferentes nombres en inglés como sentiment analysis, opinion mining, brand monitoring, buzz monitoring, online anthropology, market influence analytics, conversation mining, online consumer intelligence, user generated content) es el proceso que determina el tono emocional que hay detrás de una palabras determinadas, si una frase contiene una opinión positiva o negativa sobre un producto, marca, institución, organización, empresa, evento o persona.

Palabras difíciles

Los sentimientos se clasifican en positivos, negativos o neutros. Sin embargo, el lenguaje natural es complejo y ambiguo por lo que enseñar a una máquina a que analice los diferentes matices gramaticales, variaciones culturales, jergas, expresiones coloquiales o a distinguir faltas de ortografía, la sinonimia o la polisemia dentro de un contexto que determina el tono de la conversación es francamente difícil. Así, por ejemplo, ante un comentario sarcástico, la máquina tomaría la frase como algo positivo en vez de algo negativo o expresiones como "LOL, OMG, estuvo geeeeeeeniaaaaaaaal" son dificilísimas de procesar.

4.5. Fase 3 -

http://www.outono.net/elentir/2013/02/02/twitter-5-formas-de-identificar-a-un-troll-y
https://www.fucsia.co/opinion/blogs/entrada-blog/
como-detectar-mentiras-internet/35060 https://www.biobiochile.cl/
noticias/2013/09/16/3-senales-para-detectar-cuando-te-mienten-en-facebook-o-whatsapp.
shtml https://www.elconfidencial.com/tecnologia/2014-02-27/
pheme-un-detector-de-mentiras-para-las-redes-sociales_94300/

4.6. Utilización de la aplicación

5. Costes del proyecto

5.1. Costes económicos

5.2. Costes temporales

6. Comentario final

6.1. Conclusiones

6.2. Objetivos

El objetivo principal de este proyecto es poder dar verosimilitud sobre lo que se lee en las redes sociales, en concreto a esas personas que quieren estar informadas de lo que sucede en cuestión de segundos y ,que para ello, no se esperan a que salga en los diarios si no que se informan y se creen todo lo que leen por las redes sociales.
También se pretende que todos aquellos informativos y diarios que buscan noticias por las redes sociales puedan saber con firmeza que aquella noticia será cierta y que no se preocupen de que puedan ser víctimas de un *troll*.

7. Líneas de futuro

Comentar:

- Continuar TFM apuestas
- Sugerencias de a que personas seguir
- Dejar de seguir a gente 'agresiva'
- Facilidad en recopilación de información
- Añadir + temas no solo deportes
- Idioma

8. Bibliografía

TODO: formato correcto iso

https://docs.mongodb.com/manual/installation/

http://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

http://www.wikilengua.org/index.php/Glosario_de_deportes

9. Anexos

9.1. Palabras clave

9.1.1. Atletismo

Wgx, jogging, maratón, maratoniano, iron man, pentatlón, pértiga, skipping, bolt starting blocs, steeple-chase, tartán, fondista, runner, water jump, relevista, marcha, marchador, jabalina, javelin, lanzador de jabalina, lanzador de disco, lanzamiento de jabalina, lanzamiento de disco, lanzamiento de martillo, salto de longitud, saltador de longitud, triple salto, salto con pértiga, pertiguista, atleta, atletismo, pentathlete, pentatleta, usai, phelps, michael phelps, ruth betia, carl lewis, yelena isinbayeva, haile gebrselassie, michael johnson, jesse owens, Hicham El Guerrouj, Kenenisa Bekele, jamaica, jamaicano, Javier Sotomayor, Paquillo Fernández, Maria Lourdes Mutola, 110m, 110m vallas, javier Gómez Noya, Ashton Eaton.

9.1.2. Baloncesto

Baloncesto, basketball, básquetbol, basquetbol, baloncestista, basket, básquet, basquet, canasta, basquetbolista, dos puntos, tres puntos, nba, acb, fiba, lakers, boston celtics, ccleveland cavaliers, san antonio spurs, golden state warrios, philadelphia 76ers, chicago bulls, toronto raptors, michael jordan, miami heat, knicks, pau gasol, Kobe Bryant, Magic Johnson, Larry bird, o'neal, saquille o'neal, allen iverson, lebron james, tim duncan, kevin garnet, stephen curry, steve nash, kevin durant, derrick rose, ricki rubio, rudy fernández, josé calderoón, marc gasol, sergio llull, felipe reyes, jorge garbajosa, fernando romay, víctor claver, final four, mate, alley loop.

9.1.3. Balonmano

balonmano, handball, hándol, Nikola Karabatić, Talant Dujshebaev, vano Balic, IHF, ahf, cahb, pathf, ohf, ehf, asobal, joan cañellas, nikola karabatic, mikkel hansen, domagoj duvnjak, laszlo nagy, timur dibirov, dean bombac, arpad sterbik, julken aginagalde, kiril lazarov, daniel narcisse, filip jícha, stawomir szmal, andreas wolff, cristina neagy, eduarda amorin, andrea lekic, alexandra do nascimiento, heidi lekic.

9.1.4. Boxeo

Boxeo, boxeen, boxing day, boxing, crochet, cuadrilátero, cup protector, noquear, noqueo, punch, punching bag, punching ball, lona, ring, ring side, sparring, peso pesa-

do, pesos pesados, peso lijero, pesos lijeros, peso mosca, peso gallo, peso wélter, peso semipesado, peso semipesados, peso medio, peso superpesado, muhammad ali, rocky, rocky marciano, emilie griffith, sugar ray robinson, joe louis, julio césar chávez, floyd, mayweather, floyd mayweather, título mundial.

9.1.5. Ciclismo

Ciclismo, ciclista, ciclistas, Eddy Merckx, Bernard Hinault, Fausto Coppi, Jacques Anquetil, Gino Bartali, Miguel Indurain, Nairo Quintana, Alberto Contador, Chris Froome, Alejandro Valverde, Peter Sagan, Vincenzo Nibali, Tom Dumoulin, Mark Cavendish, Greg Van Avermaet, bici, bicis, ruedas, rueda, bicicleta, bicicletas, vuelta españa, vuelta de españa, tour de francia, france tour, tour francia, Giro d´Italia, Giro de Italia, Giro Italia, 101 kilómetros de Ronda, Volta a Portugal, dos ruedas.

9.1.6. Fútbol

Futbol, fútbol, football, soccer, futbolista, larguero, paradinha, cola de vaca, A.F.A, F.E.F, fifa, F.I.F.A, aneff, uefa, balonpie, balonpié, bundesliga, Premier League, liga bbva, liga santander, liga de campeones, champions legue, panenka, hat trick, hat-trick, Diablos rojos, Los Blues, boixos nois, naranja mecánica, submarino amarillo, eje de la zaga, tarjeta amarilla, tarjeta roja, balon de oro, balón de oro, bota de oro, entre los tres palos, cabecear, línea de cal, canaletas, cartulina, casión, espinillera, espaldinha, goleador, efiCacia goleadora, equipo ascensor, folha seca, gambeta, killer, keeper, media luna, coutinho, var, pichichi, proroga, minuto 92, minuto 93, tiempo de descuentO, puntapié, racha goleadora, messi, lionel messi, ronaldo, maradona, pele, pelé, ronaldinho, diego armando maradona, zidane, zinedine zidane, di stefano, cruyff, beckenbauer, gErard pique, busquets, cr7, d10s, arbeloa, alvaro arbeloa, roncero, neymar, de gea, bellerín, ramos, sergio ramos, pogba, arturo vidal, hazard.

9.1.7. Fútbol Americano

NFL, futbol americano, fútnol americano, american football, super bowl, quarterback, qb, touchdown, touch down, td, yac, Barry Sanders, Jerry Rice, Walter Payton, Joe Montana, Peyton Manning, Lawrence Taylor, Johnny Unitas, Reggie White, Jim Brown, Joe Greene, ncaa.

9.1.8. Golf

Hoyo, golf, golfista, hole in one, hoyo en uno, dougle eagle, madera 1, approach, backspin, tiger woods, seve ballesteros, jon rahm, roy mcllroy, jordan spierh, josé maría olazábal, dustin johnson, masters de augusta, the us masters, pga, bola wound de núcleo líquido, bola wound con el núcleo sólido, bola múlticapas, chaqueta verde.

9.1.9. Judo

Chui, dan, doyo, hayime, ipon, ippon, judo, kachi, cinturón negro, koka, kuzushi, matel, oseakomi, randori, sono mama, sore made, suri-ashi, tatami, toketa, tori, uke, ukemi, waza ar, yudo, yudogui, quimono, yudoca, judoca, yuko,yamashita yoshiaki, isogai hajime, suichi nagaoka, kelmendi majilinda, kuziutina, miranda erika, nakamura misato, shishime, siuffrida, cohen gili, krasniq, ma yingnan, kata.

9.1.10. Motociclismo

Motorista, motociclista, motociclismo, motocicleta, gp, supErbike, supersport, superstrock, motocross, trial, enduro, supermoto, rally raid, loasail, marc márquez, Andrea Dovizioso, Cal Crutchlow, lorenzo, Dovizioso, Johann Zarco, Maverick Viñales, Jack Miller, Danilo Petrucci, Valentino Rossi, rossi, Alex Rins, Andrea Iannone, Iannone, Esteve Rabat, Dani Pedrosa, Álvaro Bautista, Jorge Lorenzo, Aleix Espargaró, nicky hayden, dos ruedas, mundial de motociclismo.

9.1.11. Rugby

Rugby, rugbi, tackle, placaje, ruck, melé espontánea, melé, maul, mêlée, scrumensayo de castigo, rugbier, talonador, rugbista, medio melé, scrum-half, Springboks, All Blacks, pumas, Les Bleus, nations cup, 6 naciones, seis naciones, Beauden Barrett, Daniel Carter, Brodie Retallick, Kieran Read, Thierry Dusautoir, Richie McCaw, Shane Williams, Bryan Habana, Jonny Wilkinson, jonah lomu, Sebatien Chabal, ma'a nonu, Juan Martin Hernandez, Kees Meeuws.

9.1.12. Tenis

Tenis, tenista, raqueta, Match ball, match point, rafael nadal, rafa nadal, roger federer, novak djokovi, andy murray, del potro, david ferrer, marin cilic, tomas berdych, serena williams, hermanas williams, fernando verdasco, feliciano lopez, david goffin, john isner, maria sharapova, simona halep, richard gasquet, carlos moyá, juan carlos ferrero,

doble falta, set en blanco, ojo de halcón, alley, atp, punto de partida, punto de rotura, open australia, roland garros, wimbledon, us open.

9.1.13. Voleibol

Voleibol, volleyball, vóleibol, volibol, voleybol, voleyball, vóley, balonvolea, fivb, Gilberto Godoy Filho, giba, Wilfrido León, Ivan Zaytsev, Mariusz Wlazly.

9.2. Palabras secundarias

9.2.1. Atletismo

Carrera, salto, flat, liso, plano, obstáculos, recorrido, record, deportista, élite, miler, pista, pistas, velocidad, lisos, 200, 1500, natación, oro, plata, bronce, curb, aceleración, acelerar, sprint, final, miller, ría, relevos, walker, cross, disco, swing, stop board, throw, lanzar, lanzamiento, martillo, peso, saltar, triplista, pole, ACLIMATACIÓN, esróbico, agilidad, arena, articulaciones, batida, carrera continua, circuito, colchoneta, cuerda, fondo, photo finish, photo-finish, jaula, intervalo, s, m, resistencia, serie, tabla, justo, juegos, olimpiada, hinvierno, campeon, campeón, 800, 5000, 5.000, 10000, 10.000, 100, triple salto, vallas, metros, 4x100, YOHAN BLAKE, WAYDE VAN NIEKERK, mundial, físico, juegos de río, juegos atenas, juegos pekín, CATERINE IBARGUEN, campeona olímpica, campeon olímpico, olimpiadas, olímpico, olímpica, jjoo, estadounidense, gandaor, heptalón, disciplina, subcampeón, prueba, distancia, campeonato, marca, record del mundo, campeonato de la PGA.

9.2.2. Baloncesto

Balon, balón, equipo, jugador, cinco, tiro, puntos, liga, olímpicos, draft, falta, control, pívot, entrenador, spurs, cancha, arbitraje, tiro a canasta, bote, defensa, falta en ataque, asistencia, bloqueo, canasta limpia, contraataque, pase de pecho, pase picado, pasos, rebote, penetración, tiro lible, tiro en suspensión, base, ala-pívot, play-off, fase final, oro, plata, bronce, drafteado, rookie, cheerleader, triple, tres, dos, más uno, más uno, zona press, marcaje, zona, zonal, lob, fly, flight, guard, escolta, conductor, campo atrás, campo atras, cesta, 10 minutos, aro,juez principal, umpire, alero, defensa al hombre, falta personal, jugada de cuatro puntos, jugada de tres puntos, salto entre dos, pivotar, pantalla, gancho, buzzer beater, cross-over, bandeja, tablero, área restringida, slam dunk, cesta limpia, tocó el aro, tapon, tapón, chapa, anotó, venda, vendaje, entró, sustituir, defensor.

9.2.3. Balonmano

Fly, vuelo, marcar, pívot, time out, falta, defensa, ala, portero, fundéu, jugadores, línea, balón, lanzar, gol, central, tiempos, guardameta, descanso, banda, goles, siete contra siete, línea de 7, línea de 9, metros, línea de limitación, línea de gol, línea discontinua, 7 metros, 9 metros, amonestado, tarjeta amarilla, exclusión, sancionado, sanción, ángulo corto, ángulo largo, finta, feint rectificado, lanzamiento, golpe franco, lanzamiento recti-

ficado, shot, fallaway, hip throw, jump shot, bloqueo blocaje, barrera, pasos, yugoslava, a dos manos, muñeca, wrist, juego, pista, anotar, equipo, cinco, cancha, contrataque, amistoso, entró, parada, sustituir, defensor.

9.2.4. Boxeo

Arbitro, golpe bajo, break, clinch, agarrar, agarre, coquille, protección, cross, cros, fighter, golpe, golpe cruzado, k.o, ko, gong, jab, gancho, kao, nocaut, knock out, knock down, speed bag, straight, superwelter, swing, uppercut, welter, cinturón, cinturón de campeon, cinturon, pugil, pugilismo, puño, puñetazo, golpear, nuca, peso, asalto, titulo, título, derribo, derribar, victoria, juez, victoria por decisión, victoria por puntos, esquina, cuerda, contra las cuerdas, campana, round, bucal, protegerse, calcertas, bata, guantes, entrenamiento, saco, todos por ko, todos por ko, derrota, puntos, caer, combate, guante, venda, vendaje, cuerdas, lanzar, lanzó.

9.2.5. Ciclismo

Coll, col, esprín, sprint, esprinter, grimpeur, malla, malla rosa, maglia, maillor, maiyót, mailót, pelouse, routier, stayer, surplace, volata, abanico, peloton, abrirse, abre, abrir, autobús, badana, biela, buje, cabra, cadencia, cazaetapas, etapa, chichoera, corona, cuernos, culotte, demarraje, desarrolo, diirección, corredor, frenos, hacer un recto, hierro, llantazo, manguito, mina, ppiano, grupo, rebufo, rodador, regular, vampiro, tirón, zapata, vuelta, tour, giro, pista, ruta, prueba modalidad, montaña, bajada, subida, carrera, obstáculos, km, distancia, manillar, rueda, rodar.

9.2.6. Fútbol

Penalti, autogol, calcio, chut, centro, falta, mano, forward, portería, porteria, patada, manotazo, chutar, tiro, disparo, arbitro, arbitrar, alirón, amago, amistoso, barça, chilena, contraataque, contragolpe, filial, vaselina, escorpion, espuela, Merengue, culé, vikingos, galo, Hooligans, tifos, tifosis, poste, madera, plancha, lateral, pivote, extremo, defensa, portero, centrocampista, volante, delantero, delantero centro, carrilero, lateral derezho, lateral izquierdo, tanto, marcar, anotar, anotó, autobús, autobus, media parte, segunda parte, primera parte, primera mitad, balón dividido, banderín, banderin, barrera, bota, brazalete, capitán, césped, clasificación, control, cuerpo técnico, defensa adelantada, derby, derechazo, zurdazo, desbordar, desborde, desmarque, despejar, despeje, triplete, doblete, elástica, derrota, escudo, estadio, eurocopa, eliminatoria, encarrilar, expulsar, fair play, farolillo rojo, fichar, gol de oro, hueco, pase, imbatido, imbatibilidad, lider, líder,

liderazgo, juego aéreo, juego limpio, jugada ensayada, jugón, jugadorazo, goal, mánager, mejora de contrato, mojar, mundialito, obús, obstrucción, omnipresente, o'rey, palco, paquete, partido, pase de la muerte, pase en profundidad, pase largo, pase raso, pay per view, peinar, pelota parada, penalty, perdonar, canaletas, neptuno, cibeles, play off, potencia, presentacion, presentación, prima, primera vuelta, segunda vuelta, puntuar, puerta grande, puerta pequeña, a puerta, a puerta vacia, quiebro, recogepelotas, rechace, reconocimiento médico, recular, rectangulo de juego, regatear, remontada, remontar, renovar, replegar, replegarse, revulsivo, rival, rivalidad, clásico, robo, rondo, rotación, rueda de premsa, saque de banda, saque de esquina, salvación, seleccionador, semifinal, final, once, sombrero, subcampeón, superioridad, suplente, suspensión cautelar, tanda de penaltis, tanda de penalties, teleespectador, bicampeón, tijera, tirarse de plancha, tiro libre, offside, esferico, el cuero, fuera de juego, pena maxima, alineación indebida, abrir la lata, abre la lata, auto pase, cambio de orientación, ariete, entrar, entró, túnel de vestuario, tunel de vestuarios, vaca sagrada, velocidad, veterano, colores, la roja, vuelta de honor, zamora, zamorana, rabona, zaga, cristiano, manos, millones, confederación, campo, madrid, barcelona, zaragoza, historia, sede, hombro, codo, antebrazo, àngulo corto, ángulo largo, palo largo, finta, golpe franco, corner, córner, catenaccio, escuadra , cancerbero, linier, fondo de la red, red, pito, marco, volea, adelantad, punto de panalty, quiniela, venda, vendaje, botas, tacos, pies, rueda el balón, sustituido, defensor.

9.2.7. Fútbol Americano

Cheerleader, cheerleaders, equipo, All Pro, Artificial turf, Assistant coaches, coach, Audible, Back, ball, balón, Beat, Blitz, carga, Blindside, Bootleg, Complete pass, Pase completo, Esquinero, Cornerback, Cut back, cut, dead ball, Depth chart, defensiva, defense, Draft, Eligible receiver, Fair catch, receptor, pañuelo, flag, casco, helmet, protecciones, football, interceptar, interceptado, pase, pass rush, presión, pateador, playbook, libro de jugadas, posesion, posesión, bowl, quarter, periodo, red flag, pañuelo rojo, arbitro, entrenador, tackle, placaje, anotar, anotación, conversion, dos puntos, West Coast Offense, Wide receiver, Wildcard, Wildcat offense, ala, ofensiva wildcat, receptor, huddle, runningback, fullback, cuarto down, hole, pases, pase completo, pase incompleto, snap, scrimmage, línea de scrimmage, balón suelto, fumble, Umpire, Head Linesman, back judge, field judge, side judge, falta personal, partido, juego, final, falta, equipo, victoria, campo, suplente, mano, manotazo, venda, vendaje, botas, tacos, entre palos, sustituir, defensor, palos, derribo, derribar, chute, chutar, lanzar, lanzó.

9.2.8. Golf

Agujero, calle, driving range, campo de prácticas, fairway, green, verde, hirba fina, búnker, arenero, arena, obstáculo, hazard, water hazard, regata, charco, rough, matorral, tee box, zona de salida, course, recorrido, link, marshal, supervisor, supervisor de juego, supervisor de pista, encargado, hole, albatross, eagle, birdie, par, doble bogey, driver, madera, hierro, híbrido, putter, chip, dogleg, drive, slice, top spin, golpe, golpe con efecto, golpe largo, torpar, putt, golpe corto, open, master, major, sergio garcia, campo, hoyos, bogey, swing, mango, tecnica, ega, hook, loft, match play, medal play, stroke play, juego por golpes, eclectic, establefor, mulligan, us open, the u.s open, guantes, spike, spikes, bola, palos.

9.2.9. Judo

Victoria, amonestado, amonestación, arbitro, falta, falta leve, árbitro, interrupción, caída, kimono, delgado angelica, técnica, cuello, maestro, rojo, azul, amarillo, judogui, grado, japonés, olímpico, piernas, caer, cayó, antebrazo, potencia, demostración, naranja, combate, oro, plata, bronce, peso, grand slam.

9.2.10. Motociclismo

Velocidad, velocidades, freno, caer, tiro, motro, distancia, raid, circuito, montmeló, piloto, vehiculos, vehiculo, asfalto, curva, 125, 250, campeonato, jerez, termas de río hondo, las américas, le mans, mugello, assen, sachsenring, brno, red bull ring, silverstone, misano, chang international circuit, motegi, phillip island, sepang, mono, casco, manillar, pnéumatico, pnéumaticos, rueda, rodar, boxes, slick, liso, neumático, neumáticos, cochera, neumáticos mixtos, neumáticos de lluvia, safety car, pit wall, pole position, pit, parrilla, parrila de salida, penalización, stop and go, team radio, paddock, motorhome, warm up, grip, chattering, wheelie, caballito, feeling, rookie, motohome, piano, pianos, salida, yamaha, suzuki, honda, pit-lane, pit lane, qualifying, entrenamientos libres, entrenamiento libe, marca.

9.2.11. Rugby

Avant, fuera de juego, offside, onside, pase forward, pase adelantado, ofensivo, defensa, mark, golpe, golpear, puntapié, franco, touch, zonas laterales, line-out, ensayo, try, free-kick, kick-off, foul, foul play, zona de ensayo, zona de marca, knock-on, adelantado, penalty, catigo, penal, drop goal, conversion goal, transformación, conversión, delantera, primera línea, pilar, pilar izquierdo, pilar derecho, prop, lock, segunda línea, hooker,

second row, tercera línea, flanker, ala, centro, línea de tres cuartos, backs, fly-half, apertura, wing, zaguero, contacto, 15, 22, balón, ovalado, infracción, infracciones, naciones, Numero 8, anotar, pasar, robar, equipo, rival, venda, vendaje, botas, tacos, entre palos, sustituir, defensor, palos, touche, ruck, derribo, derribar, chute, chutar, lanzar, lanzó.

9.2.12. Tenis

Ball, pelota, break, saque, servicio, cannon ball, cañonazo, cortada, efecto, bola con efecto, deuce, drive, juego, game, spin, recogepelotas, liftado, lob, globo, bola rápida, golpe liso, net, red, passing shot, set ball, set point, punto, smash, esmach, esmachar, swing, balanceo, timing, individual, dobles, hierba, tierra, dejada, stop volley, court, pist, indoor, iguales, 30, 15, 40, falta, peloteo, individuales, bote, juez, master, mejor de tres, mejor de cinco, volea, juez de línea, dura, ladrillo, polvo, césped, in, out, tie-breack, ace, reves, set, tie-break, batida, grand slam, tierra batida, final, puntos, eliminatoria, partido, venda, vendaje, sube.

9.2.13. Voleibol

Ball in, ball out, bolea, lok out, power service, servicio, esmachado, mate, set average, tie-break, Adelantado, alcance, árbitro, auxiliar, ataque, bandas, bola muerta, zona cambio, rotación, zaguera, anotaciones, anotación, bancos, zona entrenador, red, fuera, diagonal, envío, engaño, pasadores, pasador, central, libre, puesto, atacadores, auxiliares, finta, formacion, invasión, mas-menos, net, quinto set, recibidor, señal, atrasado, varillas,voleo, zaga, k-i, k-ii, punto, ser decisivo, derrota, final, remontar, potencia, vuelo, pista, anotar, anotación, sustituir, defensor.

9.3. Palabras excluyentes

9.3.1. Atletismo

Neumático, portería, pelota, balón, placaje, boxes.

9.3.2. Baloncesto

Raqueta, guantes, rueda, portería, volea, césped, boxes, hierro, madera, chute, chutar, neumático.

9.3.3. Balonmano

Raqueta, canasta, césped, boxes, chute, chutar, neumático.

9.3.4. Boxeo

Neumático, ruedan, rueda, portería, portero, canasta, red, césped, placaje, boxes, madera, chute, chutar.

9.3.5. Ciclismo

Lanzó, pelota, jugador, jugadores, porteria, canasta,volea, balón, lona, placaje, árbitro, madera, chute, chutar.

9.3.6. Fútbol

Raqueta, guantes, boxes, neumático.

9.3.7. Fútbol Americano

Raqueta, red, rueda, portería, portero, canasta, boxes, neumático.

9.3.8. Golf

Pelota, neumático, balón, portería, canasta, portero, placaje, raqueta, boxes, madera, equipo, balón, chute, chutar.

9.3.9. Judo

Neumático, pelota, balón, raqueta, boxes, chute, chutar.

9.3.10. Motociclismo

Portero, lanzó, pelota, jugador, jugadores, portería, canasta, volea, balón, lona, placaje, árbitro, Juegos olímpicos, hierro, chute, chutar.

9.3.11. Rugby

Raqueta, red, rueda, portería, canasta, boxes.

9.3.12. Tenis

Guantes, rueda, portería, canasta, placaje, boxes, madera, chute, chutar, portero, pneumático.

9.3.13. Voleibol

Raqueta, rueda, portería, canasta, boXes, portero, chute, chutar, madera.

9.4. Palabras vacías

el él ésta éstas éste éstos última últimas último últimos a añadió aún actualmente adelante además afirmó agregó ahí ahora al algún algo alguna algunas alguno algunos alrededor ambos ante anterior antes apenas aproximadamente aquí así aseguró aunque ayer bajo bien buen buena buenas bueno buenos cómo cada casi cerca cierto cinco comentó como con conocer consideró considera contra cosas creo cual cuales cualquier cuando cuanto cuatro cuenta da dado dan dar de debe deben debido decir dejó del demás dentro desde después dice dicen dicho dieron diferente diferentes dijeron dijo dio donde dos durante e ejemplo el ella ella ello ellos embargo en encuentra entonces entre era eran es esa esas ese eso esos está están esta estaba estaban estamos estar estará estas este esto estos estoy estuvo ex existe existen explicó expresó fin fue fuera fueron gran grandes ha había habían haber hAbrá hace hacen hacer hacerlo hacia haciendo han hasta hay haya he hecho hemos hicieron hizo hoy hubo igual incluso indicó informó junto la lado las le les llegó lleva llevar lo los luego lugar más manera manifestó mayor me mediante mejor mencionó menos mi mientras misma mismas mismo mismos momento mucha muchas mucho muchos muy nada nadie ni ningún ninguna ningunas ninguno ningunos no nos nosotras nosotros nuestra nuestras nuestro nuestros nueva nuevas nuevo nuevos nunca o ocho otra otras otro otros para parece parte partir pasada pasado pero pesar poca pocas poco pocos podemos podrá podrán podría podrían poner por porque posible próximo próximos primer primera primero primeros principalmente propia propias propio propios pudo pueda puede pueden pues qué que quedó queremos quién quien quienes quiere realizó realizado realizar respecto sí sólo se señaló sea sean según segunda segundo seis ser será serán sería si sido siempre siendo siete sigue siguiente sin sino sobre sola solamente solas solo solos son su sus tal también tampoco tan tanto tenía tendrá tendrán tenemos tener tenga tengo tenido tercera tiene tienen toda todas todavía todo todos total tras trata través tres tuvo un una unas uno unos usted va vamos van varias varios veces ver vez y ya yo la en A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z.

9.5. Palabras en análisis del sentimiento

9.5.1. Alegría

Gozo, contento, deleite, diversión, placer, gratificación, satisfacción, euforia, éxtasis, emoción, jovialidad, felicidad, euforia, contento, triunfo, fascinación, dicha, alegría, júbilo, entusiasmo, estimulo, impaciencia, alivio, alborozo, deleite, jolgorio, satisfacción, esperanza, regocijo, humor, gozo, brío, agradecimiento, excitación, orgullo, embeleso, emocionado, sensual, energético, alegre, creativo, esperanzado, atrevido, estimulante, divertido, optimista, agrado, aleluya.

9.5.2. Amor

Adoración. atracción. sentimentalismo. añoranza. afecto. cuidado. deseo. amor. ternura. pasión. cariño. compasión. capricho. simpatía.

9.5.3. Enfado

Rabia, enojo, resentimiento, furia, exasperación, indignación, animosidad, irratibilidad, irritación, hostilidad, odio, violencia, malhumor, amargura, venganza, desprecio, exasperación, furia, odio, desagrado, envidia, inquietud, frustración, cólera, aversión, resentimiento, fastidio, enfado, hostilidad, menosprecio, repugnancia, aspereza, ira, violencia, rencor, distante, sarcástico, frustrado, celoso, escéptico, herir, hostil, egoísta, odioso.

9.5.4. Miedo

Angustia, alarma, aprensión, fobia, pánico, preocupación, desasosiego, incertidumbre, ansiedad, inquietud, terror, nerviosismo, aficcón, shock, pánico, tensión, pavor, miedo, histerismo, desasosiego, susto, humillación, preocupación, horror, ansiedad.

9.5.5. Sorpresa

Asombro, sorpresa, pasmo, confuso, desconcertado, chasco, conmoción, estupefacción, estupor, exclamación, extrañeza, impresión, pasmo, sobresalto, susto.

9.5.6. Tristeza

Auto compasión, soledad, desaliento, melancolía, depresión, aflicción, pena, desconsuelo, pesimismo, desesperación, tormento, pesimismo, pesar, decepción, remordimiento, rechazo, bochorno, sufrimiento, congoja, disgusto, alineación, humillación, depresión, suplicio, culpa, aislamiento, derrota, insulto, tristeza, melancolía, vergüenza, abandono,

desánimo, lástima, desesperación, infelicidad, dolor, desaliento, arrepentimiento, soledad, inseguridad, condolencia, pesadez, arrepentido, estúpido, inferior, aislado, apático, soñoliento, gulty, avergonzado, Deprimido, solitario, aburrido, cansado.