Escuela Politécnica Superior

Análisis y propuestas de técnicas de Procesamiento de Lenguaje Natural para redactar informes meteorológicos

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autor:
Rafael Soria Diez

Tutor/es:
Elena Lloret Pastor

Universitat d'Alacant
Universidad de Alicante

Mayo 2019

Justificación y objetivos

El motivo principal de elegir este proyecto fue la familiarización sobre la tecnología de *big data*. Ya que es una tecnología que está en pleno auge y me parece muy interesante. Yo ya había realizado un curso de verano en esta universidad sobre este concepto: «Big Data: fundamentos tecnológicos y aplicaciones practicas», sin embargo quería obtener más conocimientos sobre este campo y realizar un caso práctico.

Por este motivo elegí este proyecto que realiza una análisis sobre diferentes textos aplicando tecnologías que están en el mercado como PLN y GLN. Además también de utilizar otra de las tecnologías que esta muy fuerte en el mercado que es *machine learning*.

Agradecimientos

A mi tutora, Elena Lloret Pastor por aconsejarme y guiarme en este proyecto.

A mis compañeros de grado y sobretodo mis amigos: Adrián Montoya Ros, Manuel Alberola Torres y Manuel Minguez Carretero por apoyarme en los momentos más difíciles del grado y por hacer este viaje más ameno.

Y sobretodo a mis padres por apoyarme siempre y darme la oportunidad de estudiar en lo que más me gusta.

Acrónimos

- **Aemet**: Agencia Estatal de Meteorología
- **MTM**: Manual de términos meteorológicos.
- **PLN**: Procesamiento de Lenguaje Natural.
- **GLN**: Generación de Lenguaje Natural.
- **TFG**: Trabajo Fin de Grado.

Índice

1. Introducción **10**

2. Marco teórico **11**
2.1. Procesamiento de lenguaje natural 11
2.1.1. Análisis lingüístico aplicaciones 12
2.2. Propuestas existentes 13

3. Objetivos **17**

4. Metodología **18**

5. Obtención del corpus **20**
5.1. Origen de datos 20
5.2. Estructura de datos 21
5.3. Obtención de datos 26
5.4. Almacenamiento de datos 27

6. Implementación del sistema **30**
6.1. Relación datos-texto 32
6.1.1. Nubosidad 32
6.1.2. Precipitación 33
6.1.3. Tormentas 37
6.1.4. Temperatura 37
6.1.5. Viento 37
6.1.6. Visibilidad 39
6.1.7. Periodo del día 39
6.2. Clasificar texto 41
6.2.1. Conjunto de entrenamiento 42
6.2.2. Obtención del modelo de clasificación 44
6.3. Análisis de categorías 49
6.3.1. Nubosidad 51
6.3.2. Precipitación 54
6.3.3. Temperatura 59
6.3.4. Viento 63
6.3.5. Periodos del día 68
6.4. Creación de reglas 68
6.4.1. Nubosidad 70
6.4.2. Precipitación 73
6.4.3. Temperatura 76
6.4.4. Viento 78

7. Evaluación **82**
7.1. Alicante 82
7.2. Tenerife 84
7.3. Madrid 85
7.4. La Rioja 87
7.5. Conclusiones 89

8. Conclusiones **91**
8.1. Conclusiones respecto al objetivo y al aprendizaje 91
8.2. Trabajo futuro 92

9. Bibliografía y referencias **93**

Índice de figuras

1. Ejemplo de análisis sintáctico 12
2. Heliograf: política 14
3. Heliograf: deportes 14
4. Noticia redactada por Quakebot 15
5. Noticia redactada por GaliWeather 15
6. Interfaz gráfica de la api 21
7. Informe a nivel de provincia 22
8. Obteción de los textos dentro de la api 25
9. Obteción de los datos dentro de la api 26
10. Estructura de base de datos de los datos de la api 28
11. Arquitectura de generación de informes meteorológicos 30
12. Cabecera de los informes 43
13. Tabla de temperaturas de los informes 43
14. Fragmento del fichero de entrenamiento de clasificador 44
15. Ejemplo de división por KFOLD 45
16. Cabecera de informe 49
17. Fragmento del fichero csv zona-fecha-frase-categoría 49
18. Nube de palabras para la categoría nubosidad 51
19. Nube de palabras para la categoría precipitación 55
20. Nube de palabras de la categoría temperatura 59
21. Nube de palabras de la categoría viento 64
22. Ejemplo de informe final 69
23. Seis primeras horas de nubosidad 70
24. Resumen de nubosidad con sus periodos de tiempo 71
25. Agrupación de nubosidad con sus periodos de tiempo 71
26. Ejemplo de construcción de frase de nubosidad 71
27. Ejemplo de formación de periodos del día 72
28. Ejemplo de elección del verbos de nubosidad 72
29. Ejemplo de frase nube 72
30. Seis primeras horas de precipitación 74
31. Seis primeras horas de probabilidad de precipitación 74
32. Seis primeras horas de probabilidad de tormenta 74
33. Seis primeras horas de nieve 74
34. Seis primeras horas de probabilidad de nieve 74
35. Frase probabilidad 75
36. Resumen de precipitación con sus periodos de tiempo 75
37. Lista ordenada de mayor a menor 75
38. Verbos que presentan viento 75
39. Verbos de disminución del viento 76
40. Lista ordenada de mayor a menor sobre el grado de precipitaciones 76

41. Máximas temperaturas y mínimas temperaturas 77
42. Verbos de la categoría temperatura 77
43. Ejemplo de frase de cambios de temperaturas 77
44. Ejemplo de frase de heladas de la categoría temperatura 77
45. Ejemplo de frase final de la categoría temperatura 77
46. Tabla de temperaturas 78
47. Seis primeras horas de dirección del viento 79
48. Seis primeras horas de velocidad del viento 79
49. Seis primeras horas de racha máxima del viento 79
50. Resumen de la categoría viento con sus periodos de tiempo 79
51. Agrupación de viento con sus periodos de tiempo 80
52. Ejemplo de frase de dirección del viento. 80
53. Verbos de la categoría viento. 81
54. Ejemplo de frase de la categoría viento 81
55. Informe original del día 1/12/2018 de Alicante 82
56. Informe generado automáticamente del día 1/12/2018 de Alicante 82
57. Informe original del día 8/1/2019 de Alicante 83
58. Informe generado automáticamente del día 8/1/2019 de Alicante 83
59. Informe original del día 1/12/2018 de Tenerife 84
60. Informe generado automaticamente del día 1/12/2018 de Tenerife 84
61. Informe origina del día 8/1/2019 de Tenerife 85
62. Informe generado automáticamente del día 8/1/2019 de Tenerife 85
63. Informe original del informe original del día 1/12/2018 de Madrid 86
64. Informe generado automáticamente del día 1/12/2018 de Madrid 86
65. Informe original del día 8/1/2019 de Madrid 87
66. Informe generado automáticamente del día 8/1/2019 de Madrid 87
67. Informe original del día 1/12/2018 de La Rioja 88
68. Informe generado automaticamente día 1/12/2018 de La Rioja . . 88
69. Informe original del día 8/1/2019 de La Rioja 89
70. Informe generado automáticamente del día 8/1/2019 de La Rioja 89

Índice de tablas

1. Definición de lenguajes y bibliotecas 19
2. Estructura del conjunto de datos de Alicante 26
3. Estructura del conjunto de datos de Madrid 26
4. Estructura del conjunto de datos de La Rioja 27
5. Estructura del conjunto de datos de Santa Cruz de Tenerife 27
6. Adjetivos que acompañan a lluvias 34
7. Adjetivos que acompañan a chubascos 34
8. Adjetivos que acompañan a nieve 35
9. Adjetivos de probabilidad de precipitación 36
10. Adjetivos de distribución espacial de precipitación 36
11. Adjetivos de evolución temporal de precipitación 36
12. Adjetivos de heladas de la categoría temperatura 37
13. Adjetivos que definen cambios de temperatura 37
14. Adjetivos de velocidad de la categoría viento 38
15. Adjetivos que acompañan a rachas del viento 39
16. Longitud media y cantidad de nubosidad 51
17. Frecuencia de sustantivos de la categoría nubosidad 52
18. Frecuencia de adjetivos de la categoría nubosidad 52
19. Frecuencia de adverbios de la categoría nubosidad 52
20. Frecuencia de verbos de la categoría nubosidad 52
21. Verbos que indican un aumento el grado de nubosidad 53
22. Verbos que indican una disminución del grado de nubosidad . . 53
23. Verbos que indican presentación de nubosidad 53
24. Longitud media y cantidad de la categoría precipitación 54
25. Frecuencia de sustantivos de la categoría precipitación 55
26. Frecuencia de adjetivos de la categoría precipitación 55
27. Frecuencia de verbos de la categoría precipitación 56
28. Frecuencia de adverbios de la categoría precipitación 56
29. Comparación del adjetivo «débil» de la categoría de precipitación 57
30. Comparación de adjetivo «intenso» de la categoría de precipitación 57
31. Vocabulario ampliado de de entrada de datos de la categoría precipitación 58
32. Longitud media y cantidad de la categoría temperatura 59
33. Frecuencia de sustantivos de la categoría temperatura 60
34. Frecuencia de adjetivos de la categoría temperatura 60
35. Frecuencia de adverbios de la categoría temperatura 60
36. Frecuencia de verbos de la categoría temperatura 60
37. Verbos que indican aumento de la temperaturaa 61
38. Verbos que indican una disminución de la temperatura 61

39. Comparación del verbo «incrementar» en la categoría de temperatura 61
40. Comparación del verbo «descender» en la categoría de temperatura 61
41. Comparación del adjetivo «extraordinario» de la categoría temperatura 62
42. Comparación del adjetivo «notable» de la categoría temperatura . 62
43. Comparación del adjetivo «ligero» de la categoría temperatura . . 62
44. Comparación del adjetivo «poco» de la categoría temperatura . . 62
45. Comparación del adjetivo «fuerte» en la categoría de temperatura 63
46. Vocabulario ampliado de entrada de datos de la categoría temperatura 63
47. Longitud media y cantidad de la categoría viento 64
48. Frecuencia de sustantivos de la categoría viento 65
49. Frecuencia de adjetivos de la categoría viento 65
50. Frecuencia de adverbios de la categoría viento 65
51. Frecuencia de verbos de la categoría viento 65
52. Verbos que indican un aumento de la categoría viento 66
53. Verbos que indican disminución de la categoría viento 66
54. Verbos que indican presentación de la categoría viento) 66
55. Verbos de dirección de la categoría viento 66
56. Comparación del adjetivo «componente» de la categoría viento . . 67
57. Comparación de adjetivo «intenso» de la categoría viento 67
58. Vocabulario ampliado de entrada de datos de la categoría viento . 68
59. Variables de entrada de la categoría precipitación 73
60. Variables de entrada de la categoría viento 78
61. Categorías del informe original del día 1/12/2018 de Alicante . . 83
62. Categorías de informe automático del día 1/12/2018 de Alicante 83
63. Categorías del informe origina del día 8/1/2019 de Alicante . . . 83
64. Categorías del informe automático del día 8/1/2019 de Alicante . 83
65. Categorías del informe original del día 1/12/2018 de Tenerife . . 84
66. Categorías del informe automático del día 1/12/2018 de Tenerife 84
67. Categorías de informe original del día 8/1/2019 de Tenerife . . . 85
68. Categorías de informe automático del día 8/1/2019 de Tenerife . 85
69. Categorías de informe original del día 1/12/2018 de Madrid . . . 86
70. Categorías de informe automático del día 1/12/2018 de Madrid . 86
71. Categorías de informe original del día 8/1/2019 de Madrid . . . 87
72. Categorías de informe automático del día 8/1/2019de Madrid . . 87
73. Categorías de informe original día 1/12/2018 de La Rioja 88
74. Categorías de informe automático día 1/12/2018 de La Rioja . . . 88
75. Categorías de informe origina del día 8/1/2019 de La Rioja 89
76. Categorías de informe automático del día 8/1/2019 de La Rioja . 89
77. Comparaciones de los resultados entre los informes 90

1. Introducción

Uno de los problemas que ha surgido con la globalización, el crecimiento de la población y el gran avance en tecnologías es la gran cantidad de datos que circulan y se generan al diario [1]. Esta gran cantidad de datos ha generado un problema a la hora de tratarlos y para solucionarlo se ha ido construyendo diferentes algoritmos para automatizarlos y transformarlos en algo que sea rápido y fácil de asimilar para el ser humano.

Uno de los sectores que reciben datos a diario incluso minuto a minuto es el sector meteorológico. En este sector a través de sensores tanto terrestres, aéreos o marítimos recogen información para analizar el tiempo y predecirlo. Tanta cantidad de datos hay que analizarla y traducirla a un lenguaje fácil para los usuarios. Por ello este proyecto se va a centrar en la creación de un informe meteorológico automático.

Este informe automático se va a basar en unas de las tecnologías que están en auge hoy en día que son los denominados robots escritores [2]. Estos robots son capaces de escribir una noticia por ellos solos a través de una cantidad de datos dados.

Por lo tanto el objetivo de este proyecto es escribir un informe automático solamente con datos de entrada y que este informe sea leíble para un ser humano.

[1] https://www.codes.com.ar/la-inmensa-cantidad-de-datos-que-se-genera-hoy-en-dia-minuto-tras-minuto
[2] https://vanguardia.com.mx/articulo/reportero-robot-del-washington-post-ha-publicado-850-articulos

2. Marco teórico

La tecnología principal que va detrás de la creación de este informe meteorológico es el procesamiento de lenguaje natural (**PLN**). En esta sección se dará una introducción a este campo explicando su funcionamiento y citando distintas aplicaciones que tiene hoy en día. Además se citará tecnologías similares a la creación de informes meteorológicos automáticos.

2.1. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural es un campo tecnológico orientado al análisis de la lingüística que usa tecnologías de computación y de inteligencia artificial. El origen de está tecnología data de la década de 1950 con la creación del Test de Turing, sin embargo, no se llevo a cabo una gran evolución hasta la década de 1980 con la llegada de la inteligencia artificial. [3]

Su principal función es la comunicación entre el ser humano y una computadora, es decir, crear distintos mecanismos para procesar y analizar datos a un lenguaje natural, ya sea hablado o escrito. Esta fase final de creación de lenguaje es denominado generación de lenguaje natural (**GLN**[4]).

A la hora de generar lenguaje normalmente se pasa por unas etapas[5] para que quede lo más natural posible :

- **Determinar contenido**: Decidir la información que estará en el texto.
- **Estructurar contenido**: La organización del texto, por ejemplo a la hora de informar sobre el tiempo meteorológico si primero se habla de temperaturas o de precipitaciones.
- **Agregación**: Agregar las distintas frases para dar naturalidad al texto.
- **Opción léxica**: Determinar que palabras aparecen en el texto.
- **Generación de expresiones de referencia**: Generar expresiones de referencia, por ejemplo para referirse a España como el sur de Europa.
- **Realización**: Agrupar todas las etapas anteriores para formar un texto coherente.

[3] https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales
[4] https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales
[5] https://en.wikipedia.org/wiki/Natural-language_generation

2.1.1. Análisis lingüístico aplicaciones

Las principales funciones del PLN se basan en los distintos análisis lingüísticos. Entre los análisis más comunes se encuentran los siguientes:

- **Análisis morfológico**: Este análisis consiste en analizar cada palabra, es decir, identificar el lema (raíz de la palabra) de cada una identificado la presencia de prefijos, sufijos, el tiempo verbal, etc...
- **Análisis sintáctico**: Es el análisis de las funciones sintácticas. Este análisis se lleva a cabo etiquetando cada palabra y analizar la combinación de estas palabras para ver si se ha formado elaboraciones gramaticales correctas. Un ejemplo de este análisis se puede encontrar en la Figura 1.

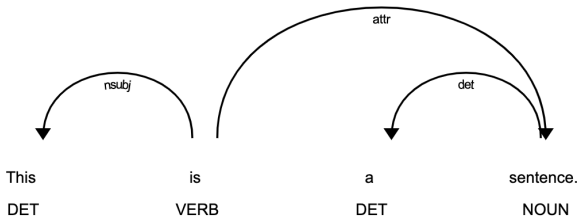

Figura 1: Ejemplo de análisis sintáctico

- **Análisis semántico**: Analiza el significado de las frases. Por ejemplo se puede encontrar frases con las mismas palabras pero con otro significado, como puede ser: «Es bueno y no es caro» y «No es bueno y es caro». Se puede ver que en estas dos frases contienen las mismas palabras pero sus significados son opuestos.
- **Análisis pragmático**: Este análisis es el encargado de analizar el contexto de las frases. Por ejemplo para ver la interpretación de su significado como puede ser la identificación de ironía.

Con estos análisis se ha desarrollado diversas aplicaciones para su uso.

- **Traducción**: La traducción automática de un lenguaje a otro. Esta traducción no se basa simplemente en traducir palabra a palabra sino que se emplea los análisis empleados anteriormente. Un ejemplo de estos traductores son Google Translate o DeepL.
- **Conversaciones automáticas**: Conversaciones entre un ser humano y una computadora.Un ejemplo de este sistema de conservaciones son los chatbot que son utilizados por múltiples empresas para resolver dudas básicas de usuarios.

- **Recuperación y extracción de la información**: Sistemas que permiten recuperar información. Esta aplicación seguramente la más utilizada por los usuarios debido a que los mayores de búsquedas de los navegadores web funcionan con esta tecnología.
- **Análisis de sentimientos**: Valorar a través de un texto el estado de ánimo del autor. Un ejemplo de uso de esta aplicación es por ejemplo valorar automáticamente los comentarios hacia una empresa o de un producto.
- **Clasificar documentos**: Clasificar documentos por categoría. De esta manera se puede hacer una separación entre distintos textos por temática. En este proyecto se utiliza esta aplicación.
- **Resúmenes automáticos**: El resumen automático es una de las aplicaciones más usadas ya que permite resumir miles de documentos para decidir por ejemplo con más rapidez si son relevantes o no.
- **Sistemas conversaciones**: Sistemas que analiza la voz del usuario y la máquina lo analiza. Un ejemplo de esta aplicación son los sistemas de voz Siri o Cortana.

2.2. Propuestas existentes

Tras ser explicado una pequeña introducción del PLN y sus aplicaciones es hora de mirar propuestas existentes a la hora de formar texto automáticamente con datos de entrada.

Entre las empresas que dan servicios de GLN se encuentra *Syllabs*[6]. Esta empresa ofrece realizar textos automáticos orientados al mercado inmobiliario. Otras de las empresas que ofrecen estos servicios pero con toda clase de temáticas son: *Arria*. [7] y *Automated Insights* [8]

En cuanto a propuestas más enfocadas a todos los usuarios se encuentra una de las propuestas que se ha hecho más famosa: un robot escritor desarrollado por el periódico *Washington Post*. Este periódico ha desarrollo un robot denominado «Heliograf» que fue capaz de escribir 850 artículos en un año [9]. Como bien indica el articulo sus principales noticias fueron redactadas en los campos de deportes y de política. Un ejemplo de los artículos sobre política y deportes:

[6] https://www.syllabs.com/es/
[7] https://www.arria.com/
[8] https://automatedinsights.com//
[9] https://www.marketingdirecto.com/anunciantes-general/medios/heliograf-robot-del-washington

Updated Apr. 8 at 11:42 a.m.

Rep. Vicente Gonzalez, a Democrat, has won another term representing Texas' 15th Congressional District, beating Tim Westley, a Republican.

With all precincts reporting, Gonzalez has 59.7 percent of the vote, followed by Westley with 38.8 percent.

Gonzalez has represented the district in the House since 2017.

Democrats have gained a majority in the House, with 235 seats in the next Congress. All of Texas's House seats have been decided; two switched hands, all going to Democrats.

The number of votes cast for Democrats in House races has increased 55.0 percent from the 2014 midterms, compared with an increase of 25.6 percent for Republicans.

The district went for Democrat Hillary Clinton in 2016 by a margin of 16.6 points. President Trump, a Republican, won Texas by nine points.

Data is from the Associated Press. This story may be updated if more information becomes available. It is powered by Heliograf, The Post's artificial intelligence system.

Figura 2: Heliograf: política

Updated Nov. 3 at 11:48 p.m.

Game Summary:
The visiting Patriot Pioneers beat the Battlefield Bobcats, 14-7, on Friday.
In the second quarter, Battlefield's Josh King was the first to put points on the board with a 11-yard touchdown reception from quarterback Graham Walker. The Pioneers' Ike Onwuka tied the score with a one-yard touchdown run.
Both teams went scoreless in the third quarter.
Entering the fourth quarter with the score tied 7-7, Patriot put the final points on the board with a 11-yard touchdown reception by Onwuka from quarterback Chris Sonnenberg.
The win improved Patriot's record to 6-2. The loss dropped Battlefield's record to 7-3.
This story may be updated if more information becomes available. It is powered by Heliograf, The Post's artificial intelligence system.

Figura 3: Heliograf: deportes

Otro de las propuestas existentes también se utiliza en el ámbito periodístico. Esta propuesta es utilizada por el periódico *Los Angeles Times*. En concreto utiliza un algoritmo denominado «Quakebot» creado por el periodística Ken Schwencke que redacta automaticamente un articulo en el caso que haya movimiento sísmicos.

Here's Monday morning's initial Quakebot report:

A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles.

According to the USGS, the epicenter was six miles from Beverly Hills, California, seven miles from Universal City, California, seven miles from Santa Monica, California and 348 miles from Sacramento, California. In the past ten days, there have been no earthquakes magnitude 3.0 and greater centered nearby.

This information comes from the USGS Earthquake Notification Service and this post was created by an algorithm written by the author.

Read more about Southern California earthquakes.

Figura 4: Noticia redactada por Quakebot

Otra propuesta similar a la planteada en este proyecto se trata de «GaliWeather»[10]. Es una aplicación que a través de sus sensores recogen datos de información meteorológica y a través de técnicas de GLN e inteligencia artificial es capaz de redactar un informe meteorológico.

	viernes, 24 de mayo			sábado, 25 de mayo			domingo, 26 de mayo			lunes, 27 de mayo		
	Mañana	Tarde	Noche	Mañana	Tarde	Noche	Mañana	Tarde	Noche	Mañana	Tarde	Noche
Estado del cielo												
Viento												
Ocurrencia lluvia	5%	5%	5%	5%	5%	5%	5%	5%	10%	40%	5%	5%
Temperatura(°C)	MIN MAX			MIN MAX			MIN MAX			MIN MAX		
Ver evolución	9°	18°		12°	21°		12°	23°		13°	21°	
Avisos Meteo										-		
Calidad aire										-		
Comentario	Se prevén cielos con nubes y claros durante los próximos días, ocasionalmente poco nubosos o despejados. En cuanto a las temperaturas, serán altas para las mínimas y normales para las máximas respecto a lo esperado para este periodo, con valores que en general ascenderán moderadamente a pesar de que oscilarán. Respecto al estado de la calidad del aire, mejorará a bueno.											

Figura 5: Noticia redactada por GaliWeather

Por último otra propuesta orientada al tiempo meteorológico es «SumTime» [11]. Esta propuesta es de habla inglesa y a través de datos meteorológicos realiza un informe explicando la evolución de estos.

[10] https://citius.usc.es/transferencia/demostradores-tecnoloxicos/GALiWeather
[11] https://www.abdn.ac.uk/ncs/departments/computing-science/sumtime-317.php

Viendo ya una lista de propuestas existentes sobre los informes meteorológicos encontramos algunas limitaciones. Por ejemplo en «SumTime» narra el tiempo meteorológico en habla inglesa, mientras que en «GaliWeather» solo se informa en la comunidad de Galicia y con sus propios sensores.

Así que que este proyecto va a intentar quitar estas limitaciones creando informes meteorológicos de habla española y recogiendo información a nivel nacional para poder realizar informes meteorológicos en todo territorio Español.

3. Objetivos

Los objetivos marcados para este proyecto son los siguientes:

1. Aprender nociones básicas sobre el procesamiento del lenguaje natural y de las tecnologías del lenguaje humano, orientadas a la generación de información.
2. Analizar y decidir las tecnologías más apropiadas para desarrollar el trabajo
3. Aprender y utilizar herramientas y librerías específicas de Procesamiento de Lenguaje Natural, así como frameworks para diseñar arquitecturas (por ejemplo, NLTK o Spacy).
4. Analizar y desarrollar un método que permita generar informes meteorológicos.
5. Evaluar el método propuesto para determinar la calidad y adecuación del informe.

4. Metodología

En esta sección se explicará la metodología llevada a cabo para el desarrollo de este proyecto y que tecnología se ha llevado a cabo.

En primer lugar la metodología que se ha llevado a cabo para la gestión del trabajo ha sido *Scrum* [12]. Este desarrollo Scrum ha consistido en reuniones periódicas con mi tutora donde definíamos que tareas había que llevar a cabo (Sprint) . Estos sprint se han estructurado en 3 partes.

Una primera parte donde se definió las tareas de la búsqueda de corpus. Un corpus es una colección de datos lingüísticos donde a través de ellos se hace un análisis para sacar conclusiones. Este obtención del corpus esta definida en la Sección 5 donde se explica todo el proceso.

Una segunda parte donde se definió las tareas de la implementación del sistema. En esta implementación entra tanto la implantación de la arquitectura para el GLN, como los pasos medios para esta implementación. En estos pasos medios entraría un clasificador de texto y diversos procesos de PLN para analizar el corpus obtenido. Este proceso de la implantación de la arquitectura se desarrolla con más detalle en la Sección 6.

Y por último una evaluación sobre el sistema automático de generación de resúmenes, donde se hace una comparación entre unos informes reales con los automáticos y se verifica la calidad de ellos. Esta última parte se corresponde a la 7

Para llevar a cabo todos estos pasos se ha hecho a través de la siguientes librerías y lenguajes de programación:

[12] `https://es.wikipedia.org/wiki/Scrum_(desarrollo_de_software)`

Tecnología		
Nombre	**Descripción General**	**Proceso**
Python	Lenguaje de programación orientado	Lenguaje base en donde se ha desarrollado todo el proyecto
Request	Biblioteca HTTP de Python que permite integraciones con sitios web.	Recogida de Datos
SQLite	Gestor de base de datos relacional	Almacenamiento de datos
Spacy	Biblioteca de PLN de Python	Análisis lingüísticos
Sklearn	Biblioteca de Python orientado al aprendizaje automático	Clasificación Texto
Pandas	Biblioteca de Python orientado al análisis de datos	Manipulación y análisis de datos

Tabla 1: Definición de lenguajes y bibliotecas

5. Obtención del corpus

En esta sección se explicará la obtención del corpus para la creación del informe meteorológico. El punto de entrada del sistema son datos meteorológicos y el resultado final es un informe meteorológico. Para hacer que esto funcione es necesario un conjunto de entrenamiento formado por datos y texto. Este texto es necesario para ver qué tipo de informe es y cómo se estructura, qué vocabulario emplea, de qué habla y la longitud de éste. En esta sección se hablará de dónde se obtienen los datos y el texto y cómo se obtienen. Los pasos a seguir:

1. Origen de datos.
2. Estructura de los datos.
3. Las decisiones realizadas para la obtención de los datos.
4. Almacenamiento de los datos.

5.1. Origen de datos

Los datos se han obtenido de la página oficial de la agencia estatal de meteorología de España: **Aemet**[13]. Está página recoge datos cada día a través de sensores y los meteorólogos realizan un informe de predicción de forma manual sobre estos.

Estos datos y sus informes se puede obtener gracias a la plataforma **Open data**[14] donde a través de un *api rest* [15] se pueden descargar los datos. Los datos de la plataforma se han obtenido del apartado **Predicción por municipios horaria** y los informes del apartado **Predicción provincial hoy**.

Para el uso de este *api rest* es necesario la obtención de un *api key*, es decir, una clave para identificar el origen de la obtención de datos. Esta key se solicita a través de un formulario indicando un correo electrónico. Un ejemplo de la interfaz gráfica de la api se muestra en la Figura 6.

[13] `https://www.aemet.es`
[14] `https://opendata.aemet.es/centrodedescargas/inicio`
[15] `https://es.wikipedia.org/wiki/Transferencia_de_Estado_Representacional`

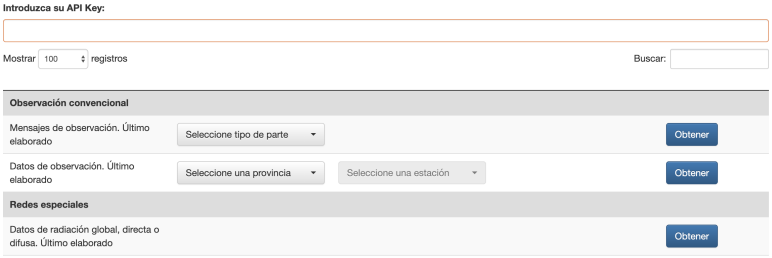

Figura 6: Interfaz gráfica de la api

5.2. Estructura de datos

Tras obtener la key ya se puede hacer uso de la api para obtener los datos y los textos.Los textos se presentan como un informe de predicción. Este informe describe el tiempo meteorológico en un día concreto,en otras palabras narra la evolución de la temperatura, viento, precipitaciones,etc... Estos informes se estructuran en tres «niveles»:

- A nivel de provincias donde narra un resumen de la provincia. Formado por datos de municipios.
- A nivel de comunidades donde narra un resumen de la comunidad. Formado por datos de provincias.
- A nivel de nacional donde narra un resumen de todo el país. Formado por datos de comunidades.

Los informes que se van a recoger se corresponden a nivel de provincia debido a que es el nivel básico del informe. Un ejemplo del informe de predicción a nivel de provincia (Figura 7):

```
AGENCIA ESTATAL DE METEOROLOGIA
PREDICCIÓN PARA LA PROVINCIA DE MADRID
DÍA 3 DE DICIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL
PREDICCIÓN VALIDA PARA EL MARTES 4


MADRID

Intervalos de nubes altas, aunque al principio del día también
se espera que esté nuboso con nubes bajas, brumas y bancos de
niebla en los valles del Tajo y Tajuña. Temperaturas mínimas en
ascenso en la Sierra y con pocos cambios en el resto; máximas en
ascenso, que será localmente notable en toda la comunidad excepto
en la Sierra. Viento flojo variable con cierto predominio de la
componente este.


TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Madrid                        6   16

Alcalá de Henares             4   18

Getafe                        4   16

Collado Villalba              7   17

Navalcarnero                  4   16

Aranjuez                      3   17
```

Figura 7: Informe a nivel de provincia

En cuanto a los datos estos tienen un formato JSON estructurados de la siguiente manera:

- Día: expresado en DD/MM/YYYY.
- Estado del cielo.
  - periodo hora a hora.
  - descripción del cielo, este tipo de dato no es numérico,sino que es un dato textual siendo sus valores:
    - Despejado.
    - Poco nuboso.
    - Intervalos nubosos/ Intervalos nubosos con lluvia/ Intervalos nubosos con nieve.
    - Nuboso / Nuboso con lluvia / Nuboso con nieve.
    - Muy nuboso / Muy nuboso con lluvia / Muy nuboso con nieve.
    - Cubierto / Cubierto con lluvia / Cubierto con nieve.
    - Nubes altas.

  Un ejemplo del fragmento JSON es el siguiente:

```
"estadoCielo" : [ {
    "value" : "11n",
    "periodo" : "00",
    "descripcion" : "Despejado"
},{
    ...
}
```

Los valores para el atributo "periodo"van del 00 hasta el 23, indicando las distintas horas del día en formato de 24 horas. Para simplificar el JSON solo se enseña el primer periodo correspondiente a la hora 00:00.

- Precipitación
  - periodo hora a hora.
  - valor de precipitación expresado en milímetro cúbicos. Este valor es un valor numérico salvo un valor textual denominado «Ip». Este valor significa que la precipitación es inapreciable (cantidad inferior a 0.1 mm^3).

```
"precipitacion" : [ {
    "precipitacion" : [ {
    "value" : "0.2",
    "periodo" : "00"
},{
    "precipitacion" : [ {
    "value" : "Ip",
    "periodo" : "01"
},{
    ...
}
```

- Probabilidad de precipitación
  - periodo en rango de 6 horas.
  - probabilidad de precipitación expresada en porcentaje. Este dato solo es numérico.

```
probPrecipitacion" : [ {
        "value" : "0",
        "periodo" : "0006"
      }, {
        "value" : "15",
        "periodo" : "0612"
```

```
        }, {
           ...
        }
```

El periodo esta escrito en formato de 4 dígitos que corresponde a un rango de horas. Por ejemplo en este caso el periodo «0006» se incumbe en el rango de horas 00:00-06:00.

- Probabilidad de tormenta
  - periodo en rango de 6 horas
  - probabilidad de tormenta expresada en porcentaje, el formato JSON es similar al «probabilidad de precipitación»
- Nieve
  - periodo hora a hora.
  - valor de nieve expresado en milímetro cúbico. Al se la nieve un tipo de precipitación, el formato JSON es similar a «precipitación» quitando el código «Ip» que en esta categoría no existe.
- Probabilidad de nieve
  - periodo en rango de 6 horas.
  - probabilidad de nieve expresada en porcentaje, el formato JSON es similar al «probabilidad de precipitación» y «probabilidad de tormenta».
- Temperatura
  - periodo hora a hora.
  - valor de temperatura expresado en grados Celsius.

```
 "temperatura" : [ {
        "value" : "14",
        "periodo" : "00"
        }, {
           ...
        }
```

- Sensación térmica
  - periodo hora a hora.
  - valor de sensación térmica expresado en grados Celsius. El formato JSON es similar a «temperatura».
- Humedad relativa

- periodo hora a hora.
- valor de humedad expresado en porcentaje.

```
"humedadRelativa" : [ {
        "value" : "54",
        "periodo" : "00"
      }, {
        ...
}
```

- Viento y racha máxima
  - periodo hora a hora del viento.
  - velocidad media del viento expresado en kilómetro/hora.
  - dirección por punto cardinal.
  - periodo hora a hora de racha máxima.
  - velocidad del racha máxima expresado en kilómetro/hora.

```
"vientoAndRachaMax" : [ {
        "direccion" : [ "N" ],
        "velocidad" : [ "5" ],
        "periodo" : "00"
      }, {
        "value" : "11",
        "periodo" : "00"
      },
```

Viendo una diferencia entre los dos tipos de información que utilizamos (datos numéricos e informes), el conjunto de datos «númerico» ya viene categorizado por su tipo, es decir, ya esta definido si pertenece a la categoría precipitación, viento, etc. Mientras que la otra parte, el conjunto de informes, no viene categorizado por ningún tipo. Por lo tanto en este conjunto de datos obtenido de Open Data se obtiene un conjunto de datos categorizados (formato JSON) y un conjunto de informes sin categorizar.

Finalmente, este conjunto de datos se generan diariamente dentro del api. Para la parte de los textos existe un historial de informes donde eligiendo un rango de días puedes obtener cualquier informe como se puede ver en la Figura 8.

Figura 8: Obteción de los textos dentro de la api

Por el contrario, la parte de los datos solo se encuentran únicamente un día en la api como se puede ver en la Figura 9. Por lo tanto, no se puede obtener los datos de una sola vez. Para ello se ha desarrollado un algoritmo que recoge el conjunto de datos y se guardan en una base de datos. Este algoritmo se ha ejecutado cada día para recoger esta información.

Figura 9: Obteción de los datos dentro de la api

5.3. Obtención de datos

Como se ha dicho anteriormente, el conjunto de datos se ha ido completando día a día consiguiendo un total de 192 días. Para acelerar este proceso y no estar tanto tiempo recogiendo datos se ha ido recogiendo información de varias provincias de manera simultánea. De esta selección se ha ido recogiendo tanto datos como informes asociados a estos datos.

Para la selección de estas provincias se ha seleccionado 4 donde el clima fuera distinto para tener más variedad de vocabulario. Por parte de ambientes fríos se han seleccionado: Madrid y La Rioja, por parte de un de ambiente seco y cálido: Alicante y por último un ambiente subtropical: Tenerife. Con el conjunto de estas 4 provincias se ha tardado un total de 48 días. Los datos recogidos por cada zona es la siguiente:

Alicante	
Informe	**Datos**
Alicante	Alicante/Alacant Orihuela Elche/Elx
	Alcoy/Alcoi

Tabla 2: Estructura del conjunto de datos de Alicante

Madrid	
Informe	**Datos**
Madrid	Madrid Alcalá de Henares Getafe Aranjuez
	Navalcarnero
	Collado Villalba

Tabla 3: Estructura del conjunto de datos de Madrid

La Rioja	
Informe	**Datos**
La Rioja	Haro Calahorra Logroño

Tabla 4: Estructura del conjunto de datos de La Rioja

Tenerife	
Informe	**Datos**
Tenerife	Santa Cruz de Tenerife

Tabla 5: Estructura del conjunto de datos de Santa Cruz de Tenerife

Las anteriores tablas muestran la estructura de las provincias y los municipios. Esta división entre las provincias y municipios ya están impuestas en la api. Por otro lado también muestra la división entre el informe y los datos. Los datos se corresponde a los distintos municipios y el informe al conjunto de estos municipios dando lugar a un informe de provincia.

Una vez explicado que estructura va a tener los datos es hora de almacenarlos.

5.4. Almacenamiento de datos

Como los datos se recogen día a día es necesario almacenarlos en una base de datos. Esta base de datos es una base de datos relacional guardada mediante el gestor de base de datos SQLite [16]. El esquema de la base de datos es el siguiente:

[16] `https://www.sqlite.org/index.html`

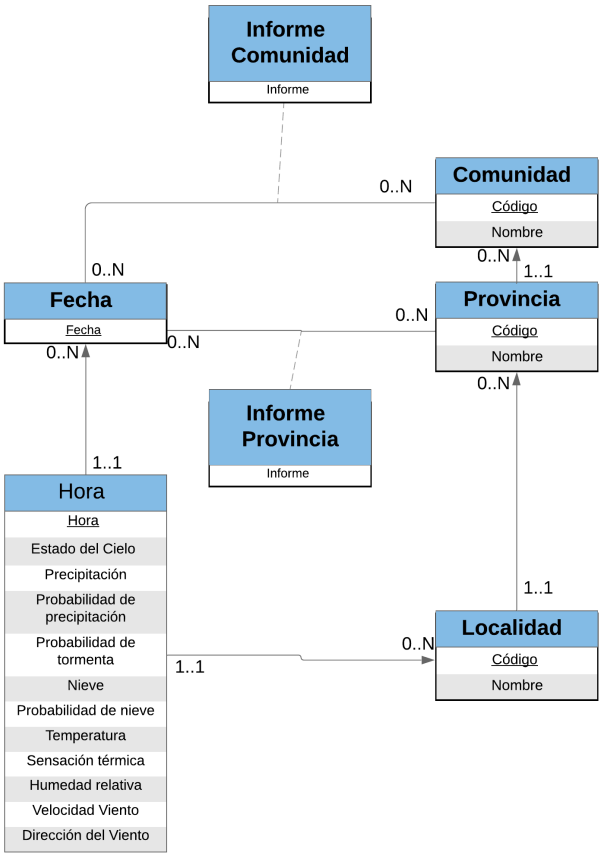

Figura 10: Estructura de base de datos de los datos de la api

Como se puede ver en el esquema de la base de datos se guarda tanto los informes y los datos correspondiente a estos informes. Los datos están guardados en la tabla «Hora» y los informes en las tablas «InformeProvincia» y «InformeComunidad». Estas se relacionan entre si con la fecha y las localidades, provincias y comunidades. El código(identificador) asociado es el mismo código que utiliza la Api para su identificación.

Una vez organizado como se guardar los datos se efectúa el almacenamiento de los datos obtenidos en los 48 días a través de un algoritmo escrito en Python 3:

1. Invocación del Api Rest a través de una llamada con la key con la librería Request.
2. Obtener la Fecha a través del JSON de datos y guardarla en la base de datos.

3. Guardar las localidades, provincias y comunidades.
4. Ordenar los datos del formato JSON de forma ordenada tal y como se estructura en la tabla «Hora».
5. Guardar los distintos informes.

6. Implementación del sistema

En este capitulo se hablará del desarrollo paso a paso de la arquitectura propuesta. Esta arquitectura, es una arquitectura híbrida que combina técnicas de *machine learning* y reglas. Esta arquitectura se muestra en la Figura 11 donde se encuentran los distintos módulos para la creación de un informe meteorológico.

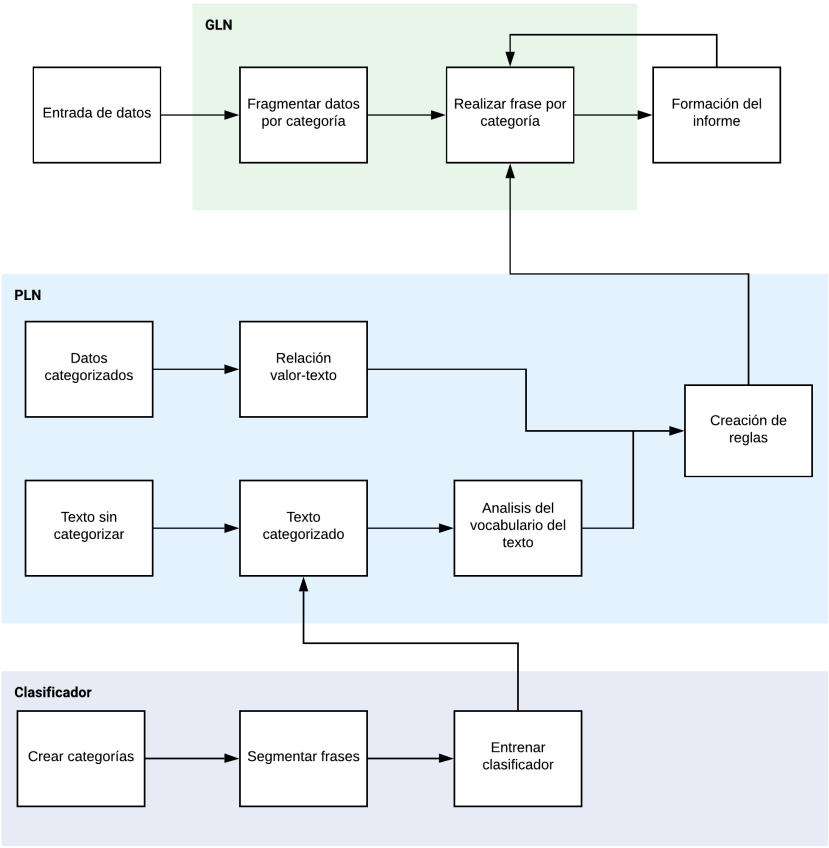

Figura 11: Arquitectura de generación de informes meteorológicos

En la figura se aprecia que la arquitectura se divide en tres partes. La primera parte se corresponde a GLN. La función de esta parte es la generación de informes meteorológicos que con datos de entrada se realiza el informe final a través de reglas. Ahora bien, para la creación de estas reglas se realiza un estudio del conjunto de datos obtenido en la Sección 5. Este estudio se corresponde a la parte de PLN.

Como se vio en la obtención de datos se recogían dos tipos de datos: uno en formato texto que se corresponde a los informes y otros datos que indican valores del tiempo y que se corresponde a los datos de entrada del sistema. Con estos dos tipos de datos se va a realizar un estudio de ambos para la creación de reglas.

Para la parte de datos numéricos se va a realizar un mapeo entre los datos y el texto, dicho de otra manera, sacar la relación entre el valor de un dato con el texto. Por ejemplo si un dato de precipitación tiene un valor de 12 averiguar si se corresponde a lluvias débiles, fuertes, etc... Para hacer este estudio e intentar tener una eficacia máxima se va a llevar a cabo a través de un manual de uso de términos meteorológicos [17] (**MTM**).

Este manual ha sido escrito por expertos meteorológicos de la propia AEMET para los usuarios. En él se encuentra un glosario de qué vocabulario hay que emplear a la hora de transmitir información del tiempo meteorológico y cómo redactarla. Por lo tanto a la hora de sacar esta relación se hará uso de este manual.

Por parte de los textos se hará un estudio del vocabulario haciendo diversos análisis de PLN. Para ello y para tener más eficacia a la hora de realizar las frases se van a clasificar las distintas frases de los textos en las categorías impuestas en los datos numéricos explicados en la Sección 5. De esta manera se puede obtener solo vocabulario de la categoría analizada y dar más cohesión a la frase. Para hacer esta clasificación se hará a través de técnicas de *machine learning* empleando clasificadores.

Por lo que esta arquitectura se puede resumir en las siguientes fases. Para la parte de GLN que es la fase final se corresponde los siguientes procesos:

1. Entrada de datos numéricos.
2. Dividir los distintos datos en categorías.
3. Creación de frases categóricas a través de reglas.
4. Construcción del informe final uniendo todas las frases por categorías.

Para la parte de PLN que se corresponde con la creación de reglas para la parte de GLN se sigue estos pasos:

1. Relacionar datos texto a través del MTM.
2. Categorizar el texto a través de machine learning.
3. Realizar un análisis del vocabulario empleado en cada categoría.

[17] `http://www.aemet.es/es/zona_portada_destacada/manual_terminos_met`

4. Creación de frases a través de reglas. Estas reglas se forman a través de la reacción entre datos y texto y el análisis del vocabulario.

A continuación se va a detallar con profundidad estos pasos.

6.1. Relación datos-texto

En esta parte se explicará la relación entre datos a texto, es decir, a que texto se transformará los datos de entrada. Esta relación como se ha dicho anteriormente se va a hacer a través del MTM. En este MTM define 6 tipos de categorías meteorológicas que debe tener cualquier informe meteorológico. Estas categorías delimitan que, dependiendo del valor de un dato se tiene que emplear un vocabulario u otro. Estas categorías coinciden con los datos obtenidos de la propia AEMET. Estas categorías impuestas por ellos son:

- Nubosidad
- Precipitación
- Tormenta
- Temperatura
- Viento
- Visibilidad

A lo largo esta sección se explicará el vocabulario que tiene cada valor por categoría. Este vocabulario esta explicado mediante tablas, en las tablas se muestra a que valor o que rango de valores será transformado los datos en texto.

6.1.1. Nubosidad

La primera categoría definida en el MTM es nubosidad. Esta categoría concuerda con el tipo de dato **estado del cielo** explicada en la Sección 5. Este dato tiene la característica propia de ser la única que contiene texto. Por consiguiente y en este caso el tipo de dato ya esta relacionado con el texto, así que a la hora de hablar de nubosidad según su valor se utilizará los siguientes términos:

- Despejado
- Poco nuboso
- Intervalos nubosos/ Intervalos nubosos con lluvia/ Intervalos nubosos con nieve
- Nuboso / Nuboso con lluvia / Nuboso con nieve

- Muy nuboso / Muy nuboso con lluvia / Muy nuboso con nieve
- Cubierto / Cubierto con lluvia / Cubierto con nieve
- Nubes altas

Por lo tanto cuando entre al sistema un tipo de dato de la categoría nubosidad se transformará directamente en estas lista de palabras.

6.1.2. Precipitación

La siguiente categoría es precipitación. En esta categoría aparecen cuatro tipos de datos definidos en la Sección 5: **precipitación**, **probabilidad de precipitación**, **nieve** y **probabilidad de nieve**.

En primer lugar se describirá la variable precipitación que según el MTM podemos definirla con tres tipos de precipitación: llovizna, lluvias y chubascos. En esta descripción se hará también mención a la definición impuesta por AEMET de estos fenomenos para distinguirlos entre si.

Llovizna

El MTM define como lloviznas: « *Precipitación cuasi uniforme compuesta exclusivamente de pequeñas gotas de agua (diámetro menor de 0,5 mm) muy numerosas, que pueden reducir la visibilidad en mayor medida que la lluvia ligera*». Según la entrada de datos de la Sección 5 esta definición concuerda con el código **Ip**. Por lo tanto cada vez que entre en el sistema un tipo de dato Ip se categorizara como llovizna.

Lluvias

El MTM define como lluvias: «*Precipitación de partículas de agua líquida en forma de gotas de diámetro mayor de 0,5 mm o más pequeñas si caen de forma dispersa*». Además proporciona los siguientes adjetivos dependiendo del grado de la lluvia:

Lluvias	
Datos	**Valor**
Débiles	Cuando su intensidad es menor o igual que 2 mm/h.
Sin adjetivar	Si su intensidad es mayor que 2 mm/h y menor o igual que 15 mm/h
Fuertes	Aquellas cuya intensidad es mayor que 15 mm/h y menor o igual que 30 mm/h.
Muy Fuertes	Si su intensidad es mayor que 30 mm/h y menor o igual que 60 mm/h.
Torrenciales	Para intensidades mayores de 60 mm/h.

Tabla 6: Adjetivos que acompañan a lluvias

Por lo tanto cuando se encuentre un tipo de dato numérico con este rango de valores se categorizará como lluvias y su correspondiente adjetivo si lo tuviera.

Chubascos

El MTM define como chubascos: «*Precipitación, frecuentemente fuerte y de corta vida, que cae desde nubes consecutivas; las gotas o partículas sólidas en los chubascos son usualmente mayores que los elementos correspondientes en otros tipos de precipitación. Los chubascos se caracterizan por su comienzo y final repentinos, generalmente por grandes y rápidos cambios de intensidad. Como sinónimo de chubasco puede usarse el término aguacero.*». Además proporciona los siguientes adjetivos dependiendo del grado de los chubascos:

Chubascos	
Datos	**Valor**
Sin adjetivar	Si su intensidad es mayor que 2 mm/h y menor o igual que 15 mm/h
Fuertes	Aquellas cuya intensidad es mayor que 15 mm/h y menor o igual que 30 mm/h.
Muy Fuertes	Si su intensidad es mayor que 30 mm/h y menor o igual que 60 mm/h.
Torrenciales	Para intensidades mayores de 60 mm/h.

Tabla 7: Adjetivos que acompañan a chubascos

Por último los chubascos son similares a las lluvias con la diferencia que no hay chubascos débiles. A la hora de diferéncialos con las lluvias se hace a través de su definición en concreto a la parte de: «se caracterizan por su comienzo y final repentino». Por lo tanto se nombrará los chubascos con su correspondiente adjetivo cuando se encuentre una precipitación suelta en una hora ya que comenzará

y se finalizará de forma repentina.

Por último y tal como se ha mencionado en la categoría nubosidad, el valor de la variable «precipitación» se transformara a texto dependiendo de la información de las tablas.

Nieve

El siguiente tipo de precipitación es nieve que se corresponde con la variable nieve obtenido de AEMET. Sus adjetivos definidos dependiendo del grado son:

Probabilidad	
Datos	**Valor**
Débiles	Los copos son normalmente pequeños y dispersos. Con viento en calma el espesor de la cubierta de nieve aumenta en una cantidad no superior a los 0,5 cm/h.
Sin adjetivar	Normalmente, consisten en copos de mayor tamaño que caen con suficiente densidad como para disminuir la visibilidad sustancialmente. La cubierta de nieve aumenta su espesor en una proporción de hasta 4 cm/h.
Fuertes:	Reduce la visibilidad a un valor bajo y aumenta la cubierta de nieve en una proporción que excede a los 4 cm/h.

Tabla 8: Adjetivos que acompañan a nieve

Por otro lado, aparte de las precipitaciones el MTM también menciona la probabilidad de éstas, sus distribuciones espaciales y su evolución temporal.

Probabilidad

Esta variable se corresponde a la variable de **probabilidad de precipitación** y **probabilidad de nieve** e indica la probabilidad de que ocurra la precipitación. Sus adjetivos dependiendo del valor:

Probabilidad	
Datos	**Valor**
Baja probabilidad	Cuando la probabilidad de que ocurra el fenómeno esté comprendida entre el 10 % y el 40 %. Con el mismo significado se podrán utilizar expresiones como “poca probabilidad”, “pequeña probabilidad”, “escasa probabilidad” o“no se descarta ...”.
Probable	Cuando la probabilidad esté entre el 40 % y el 70 %.
Sin adjetivar	No se hará mención de ningún término de probabilidad cuando esta sea mayor del 70 %.

Tabla 9: Adjetivos de probabilidad de precipitación

Distribución espacial

La distribución espacial es el porcentaje de a qué zonas se da el fenómeno de precipitación, es decir, si una provincia esta formada por 5 municipios, se define a cuántos municipios abarca el fenómeno de precipitación. Para describirlo define los siguientes adjetivos:

Distribución Espacial	
Datos	**Valor**
Aisladas o dispersas	Cuando dicho porcentaje sea inferior al 30 %.
Sin adjetivar	Cuando el porcentaje esté entre el 30 % y el 60 %.
Generalizadas	Cuando sea mayor del 60 %.

Tabla 10: Adjetivos de distribución espacial de precipitación

Evolución Temporal

La evolución temporal es el porcentaje de la cantidad de tiempo en que se produce el fenómeno de precipitación, en nuestro caso es 24 horas. Para describirlo define los siguientes adjetivos

Evolución Temporal	
Datos	**Valor**
Ocasionales	Para una duración inferior al 30 % del periodo de predicción.
Persistentes o continuas	Cuando el porcentaje esté entre el 30 % y el 60 %.
Intermitentes	Si se produce de manera casi regular, con interrupciones durante cortos intervalos de tiempo. Es evidente que cuando las precipitaciones se pueden calificar de intermitentes, la duración total del fenómeno será de aproximadamente el 50 %.

Tabla 11: Adjetivos de evolución temporal de precipitación

6.1.3. Tormentas

Esta categoría esta muy ligada a la categoría precipitación y pertenece al tipo de dato **probabilidad de tormenta**. El MTM define los mismos adjetivos que tiene probabilidad de precipitación y probabilidad de nieve definidos en la tabla 9.

6.1.4. Temperatura

La categoría de temperatura se asocia al tipo de dato **temperatura**. En esta categoría indica que a la hora de hablar de temperatura hay que reconocer dos parámetros. El primer parámetro es en caso que el valor sea inferior a 0°C. Cuando esto ocurra hay que definirlo como «heladas». Estas heladas se puede adjetivar de la siguiente manera:

Heladas	
Datos	**Valor**
Débiles	Cuando la temperatura esté entre 0°C y -4°C.
Sin adjetivar	Cuando la temperatura esté entre –4°C y –8°C .
Fuertes:	Cuando la temperatura sea inferior a –8°C.

Tabla 12: Adjetivos de heladas de la categoría temperatura

El otro parámetro corresponde a la evolución de la temperatura respecto al día anterior, es decir, dependiendo de si ha habido cambios en las máximas y mínimas se añadirán el siguiente vocabulario:

Cambios de Temperatura	
Datos	**Valor**
Pocos cambios o cambios ligeros	Variaciones de 2°C
Aumento/descenso	Variaciones de entre 3°C a 5°C.
Aumento/descenso notable	Variaciones de entre 6°C a 10°C.
Aumento/descenso extraordinario	Variaciones superiores a 10°C.

Tabla 13: Adjetivos que definen cambios de temperatura

6.1.5. Viento

Esta categoría tiene ligada las categorías de **dirección del viento**, **velocidad del viento** y **racha máxima**. A la hora de describir la dirección del viento, este tipo de variable igual que ocurre con la variable «estado del cielo» también tiene su valor en texto. En consecuencia los valores posibles son:

Dirección

- Norte.
- Noreste.
- Este.
- Sureste.
- Sur.
- Suroeste.
- Oeste.
- Noroeste.

Además de indicar la dirección en puntos cardinales señala que se puede agrupar estos puntos en el caso que la dirección del viento provenga de varias direcciones. Estas agrupaciones las hace a través de componentes:

- **Componente Norte**: Norte, Noreste y Noroeste.
- **Componente Sur**: Sur, Sureste y Suroeste.
- **Componente Este**: Este, Sureste y Noreste.
- **Componente Oeste**: Oeste, Noroeste y Suroeste.

Por lo que si llega al sistema direcciones de viento de dirección norte, noreste y noroeste se transformará directamente a «componente norte».

Velocidad

En cuanto a la velocidad para definir el grado del viento se puede añadir los siguientes adjetivos:

Velocidad	
Datos	**Valor**
Calma	Velocidad media menor o igual a 5 km/h.
Flojos	velocidad media entre 6 km/h y 20 km/h.
Sin adjetivar	Velocidad media entre 21 km/h y 40 km/h.
Fuertes	Velocidad media entre 41 km/h y 70 km/h.
Muy Fuertes	Velocidad media entre 71 km/h y 120 km/h.
Huracanados	Velocidad media mayor que 120 km/h.

Tabla 14: Adjetivos de velocidad de la categoría viento

Rachas

Por último las rachas según el MTM: solo se pueden citar cuando superen los 70 km/h y se puede adjetizar de la siguiente manera:

Rachas	
Datos	**Valor**
Muy Fuertes	Cuando superen o vayan a superar los 70 km/h.
Huracanados	Para más de 120 km/h.

Tabla 15: Adjetivos que acompañan a rachas del viento

6.1.6. Visibilidad

La última categoría es visibilidad, esta categoría según el MTM se define a través de dos variables. Una primera variable que se corresponde al tipo de dato **Humedad relativa** y otra variable que seria visibilidad. Esta última variable no se encuentra en el conjuntos de datos Open Data puesto que la define como la variable más dificil de categorizar puesto que depende de muchos factores. Por este motivo no es posible de realizar una frase sobre esta categoría.

6.1.7. Periodo del día

Por último y aunque esta parte no pertenece a ninguna categoría, el MTM menciona que cuando hay un cambio de fenómeno se puede mencionar el periodo del día en el que ocurre. Estos periodos del día las define según la salida y la puesta del sol, puesto que no hay una salida del sol y puesta del sol concreta debido a que depende de la estación del año y otros factores he decidido poner una hora de inicio y fin aproximada.

Madrugada

El MTM define «madrugada» como: « *Horas que siguen a la medianoche hasta la salida del sol*».Los datos definidos han sido:

- **medianoche**: 12:00
- **salida del sol**: 7:00

Día

El MTM define «día» como: «*Tiempo que media entre la salida y la puesta del sol*». Los datos definidos han sido:

- **Salida del sol**: 8:00
- **Puesta del sol**: 19:00

Mañana

El MTM define «mañana» como: «*Tiempo que transcurre desde que amanece hasta el mediodía*». Los datos definidos han sido:

- **Amanece**: 8:00
- **Mediodía**: 12:00

Mediodía

El MTM define «mediodía» como: « *Momento en que está el sol en el punto más alto de su elevación sobre el horizonte*». El dato definido ha sido **mediodía** con 12:00.

Horas centrales de la mañana

El MTM define «horas centrales de la mañana como»: « *Intervalo de horas en el que el sol está más alto sobre el horizonte. Se puede considerar un intervalo de más/menos 2 horas alrededor del mediodía*». El dato definido ha sido **horas centrales de la mañana** con un valor entre las 10:00-11:00 y 13:00-14:00.

Tarde

El MTM define «tarde» como: « Tiempo entre el mediodía y el anochecer (puesta del sol).». Los datos definidos han sido:

- **Mediodía**: 12:00
- **Puesta del sol**: 19:00

Noche

El MTM define «noche» como: « *Tiempo que media entre la puesta y la salida del sol*». Este dato ha sido el mas complicado de definir debido a que el conjunto de datos los días están definidos entre las horas de 00:00-23:00, por lo tanto no podría definir la salida del sol ya que contaría como el día siguiente. Por lo tanto el fin de este periodo se corresponde al fin del día:

Los datos definidos han sido:

- **Inicio**: 20:00
- **Fin**: 23:00

6.2. Clasificar texto

Una vez obtenido qué texto se obtendrá de los datos de entrada es hora de analizar el vocabulario de los informes recogidos por la base de datos. Recordando la Sección 6, el primer paso para realizar este análisis es clasificar las frases de los textos en categorías.

Por lo tanto el siguiente paso es categorizar los informes a las categorías descritas en el MTM mencionadas en la sección anterior. Se va a categorizar a las categorías descritas en la sección anterior ya que esas categorías son las que tienen que están en el informe y por lo tanto a la hora de completar una frase con una categoría concreta tiene que ser con vocabulario de esa categoría.

Para categorizar se va hacer una fragmentación de frases de los informes. Una vez fragmentado se va a categorizar de forma manual. Con esta fragmentación de frases asociado a su categoría ya se puede hacer un análisis del vocabulario de cada categoría.

Como este proceso es tedioso se va a automatizar utilizando *machine learning* para desarrollar un clasificador. Esta automatización es necesario debido a que en el caso de necesitar más informes para sacar más vocabulario o más características de los informes ya no es necesario categorizarlo a mano. Además este clasificador se utilizará en un futuro para la comparación de los informes (Sección 7) realizados de forma manual por los meteorólogos y los informes automáticos.

Para la creación de este clasificador el primer paso es sacar un dataset para entrenar. Los pasos a seguir a la hora de realizar un clasificador:

1. Creación de un dataset para entrenar.
2. Entrenar un clasificador para automatizar este proceso y obtener un modelo asociado que nos permita asociar las categorías automáticamente.

6.2.1. Conjunto de entrenamiento

Para conseguir que el clasificador funcione primero hay que construir un conjunto de entrenamiento. Para crear este conjunto de entrenamiento se va a seguir los siguientes pasos:

1. Crear Categorías.
2. Eliminar información no concluyente.
3. Fragmentar frases.
4. Tokenizar oraciones.
5. Análisis de las categorías de las palabras (Postagging).
6. Lematizar palabras.
7. Guardar frases lematizadas en un fichero de tipo .csv
8. Clasificar oraciones.

El primer paso entonces es la creación de categorías que como se había dicho en la Sección 6.1 las 6 categorías impuestas por el MTM son:

1. Nubosidad
2. Visibilidad
3. Precipitación
4. Tormentas
5. Temperatura
6. Viento

Sin embargo, como la categoría precipitación y tormentas van muy ligadas se ha decidido fusionarla en una. Además, se ha añadido la etiqueta «otros» por si a la hora de fragmentar existen frases sueltas que no pertenecen a ninguna categoría. Por lo que las categorías impuestas a clasificar son las siguientes:

1. Nubosidad
2. Visibilidad
3. Precipitación
4. Temperatura

5. Viento
6. Otros

Con las categorías impuestas, el siguiente paso es la fragmentación de los informes para obtener sus frases. Sin embargo, antes de realizar esta fragmentación se ha eliminado la cabecera (Figura 12) y la tabla de temperaturas incorporada en todos los informes (Figura 13). Ya que esta información es irrelevante a la hora de entrenar el clasificador ya que son frases que no pertenece a ninguna categoría y puede influir en la precisión del clasificador.

```
AGENCIA ESTATAL DE METEOROLOGÍA
PREDICCIÓN PARA LA PROVINCIA DE MADRID
DÍA 5 DE DICIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL
PREDICCIÓN VALIDA PARA EL JUEVES 6
```

Figura 12: Cabecera de los informes

```
TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):
Madrid                      5   15
Alcalá de Henares           1   16
Getafe                      5   15
Collado Villalba            3   16
Navalcarnero                5   16
Aranjuez                    3   15
```

Figura 13: Tabla de temperaturas de los informes

Una vez eliminado estos dos componentes es hora de fragmentar los informes en sus diferentes frases. Esta fragmentación se ha llevado a cabo mediante procesos de PLN. Una vez fragmentada las frases se ha llevado a cabo el proceso de **tokenización** que consiste en separar cada palabra del texto, sea una palabra, espacio o carácter. Una vez tokenizado se ha llevado a cabo el proceso de **postagging**.

El postagging es la etiquetación morfoléxica [18]. Esto se ha llevado a cabo porque de cada oración solo me quedo con los elementos esenciales a la hora de clasificarlas. Esta información son los sustantivos, adjetivos, verbos y adverbios dejando de lado preposiciones, conjunciones y demás elementos que comparten todas las oraciones y pueda empeorar la precisión del clasificador tal y como ocurrían con las figuras de tablas de temperaturas y la cabecera.

Por último, el último proceso de PLN es la **lematización** que consiste en sacar el lema de una palabra, es decir, la raiz de una palabra. Por ejemplo de «casas» su lema sería «casa». Esto se ha realizado para no tener que repetir las mismas palabras consiguiendo una mayor precisión a la hora de clasificar. Una vez realizado todos estos pasos se guarda las oraciones lematizadas en un fichero .csv.

[18] `http://roble.pntic.mec.es/~msanto1/lengua/1anamorf.htm`

Una vez obtenidas las frases lematizadas es hora de etiquetarlas. Las frases segmentadas da una cantidad total de 820. Esta etiquetación de todas las frases se ha efectuado de forma manual y se ha revisado un total de 3 veces en periodos distintos de tiempo para comprobar que el resultado era el mismo. Un ejemplo del fichero .csv con las oraciones lemmatizadas y clasificadas se encuentra en la figura 14

```
frase,clase
tarde viento flojo variable,viento
viento sur suroeste moderado fuerte debilitándose largo día,viento
brisas costas norte noroeste,viento
nuboso despejado,nubes
Baja probabilidad calima ligera primera mitad día,visibilidad
Temperaturas cambios ligero descenso,temperatura
Viento componente costas sureste medianías zonas altas moderado disminuyendo flojo final,viento
costa oeste,otros
Viento sureste intervalos fuerte cumbres vertientes nordeste incluida área metropolitana vertientes suroeste medianías nordeste rachas llegar
costas,viento
Temperaturas ascenso,temperatura
Viento general flojo componente norte madrugada girando componente oeste mañana predominando brisas costas,viento
cumbres centrales viento noroeste intervalos fuerte,viento
nuboso despejado general intervalos nubosos mañana norte noreste,nubes
Temperaturas cambios,temperatura
```

Figura 14: Fragmento del fichero de entrenamiento de clasificador

6.2.2. Obtención del modelo de clasificación

Obtenido ya el conjunto de entrenamiento es hora de entrenar el clasificador. El clasificador se ha hecho a través de algoritmos de machine learning de la librería «scikit-learn» [19]. Para obtener el mejor resultado de acierto se ha experimentado con dos dos clasificadores uno de tipo SVM [20] y otro de tipo Bayes[21]. Para comparar los rendimientos se ha calculado las métricas: **precisión**, **combertura**, **medida F** y **exactitud**. Antes de explicar el significado de estas métricas se va a explicar los valores de: **verdadero positivo**, **verdadero negativo**, **falso negativo** y **falso positivo**, ya que estos valores definen las métricas anteriores.

- **Verdadero positivo (TP)**: valores positivos pronosticados correctamente, es decir, si la etiquetación de un objeto es **positiva** y la predicción de un objeto es **positiva**.
- **Verdadero negativo (TN)** : valores negativos pronosticados correctamente, es decir, si la etiquetación de un objeto es **negativa** y la predicción de un objeto es **negativa**.
- **Falso positivo (FP)**: cuando la etiquetación de un objeto es **negativa** y la predicción de un objeto es **positiva**.
- **Falso negativo (FN)**: cuando la etiquetación de un objeto es **positiva** y la predicción de un objeto es **negativa**.

Una vez explicado estos valores, voy a explicar las métricas:

[19] https://scikit-learn.org/stable/dePython.
[20] https://es.wikipedia.org/wiki/Maquinas_de_vectores_de_soporte
[21] https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo

- **Precisión**: es la relación entre las relaciones positivas y el total de observaciones positivas recibidas.

$$Precisión = \frac{TP}{TP + FP} \tag{1}$$

- **Cobertura**: es la relación entre las observaciones positivas pronosticadas correctamente y todas las observaciones en el objeto.

$$Cobertura = \frac{TP}{TP + FN} \tag{2}$$

- **Medida F**: es el promedio de cobertura y precisión junto a los falsos positivos y falsos negativos

$$MedidaF = (1 + \beta^2) + \frac{Precisión \times Cobertura}{(\beta^2)TP + FN} \tag{3}$$

- **Exactitud**: es el porcentaje de acierto de predecir.

$$Exactitud = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

A la hora de elegir la proporción del conjunto de entrenamiento y el conjunto de test, se ha decidido realizar la técnica llamada de **cross validation** concretamente **k-fold**.

Esta técnica divide el conjunto total de datos en proporciones iguales, en cada iteración coge un porción de entrenamiento y otro de test y así hasta el fin de las iteraciones, pasando por todos los conjuntos de datos posibles. Al final el porcentaje de acierto final es la media de todas las iteraciones:

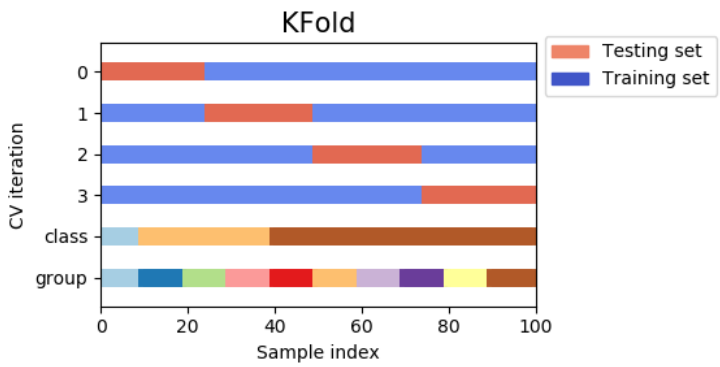

Figura 15: Ejemplo de división por KFOLD

En este proyecto los entrenamientos se han llevado a cabo un total de 10 veces.

SVM

SVM denominado **máquinas de vectores de soporte** en castellano, son un conjunto de algoritmos de *machine learning* tanto en el ámbito de clasificación como el ámbito de regresión. Este algoritmo es capaz de predecir y etiquetar unas serie de características a través de un modelo de datos. Su funcionamiento se basa en la creación de distintos hiperplanos en un espacio que clasifica las clases en dos espacios dimensionales (valores positivos o negativos explicado anteriormente). La función del hiperplano es dividir estas dos clases, y según tan cerca este de una o otra la clasificación, sera de ese tipo.

El tipo de SVM realizado es con Stochastic Gradient Descent [22] que esta orientado hacia la clasificación de texto y procesamiento de lenguaje natural.

Para la creación del SVM en primer lugar se ha creado un pipeline, un pipeline es una herramienta que proporciona scikit-learn que simplifica el desarrollo del clasificador. Su función es aplicar una serie de transformaciones de los datos y que tipo de clasificador de machine learning se va aplicar. El pipeline utilizado en este clasificador es el siguiente:

[22] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html`

```
text_clf = Pipeline([
('vect', CountVectorizer()),
('tfidf', TfidfTransformer()),
('clf', SGDClassifier(loss='hinge', penalty='l2',
                      alpha=1e-3, random_state=42,
                      max_iter=5, tol=None)),
])
```

- **vect**: Se inicializa con countVectorize() que hace el conteo de frecuencia de texto, es decir, realiza un *Bag of words*.[23] que esta orientado hacia la clasificación de texto y procesamiento de lenguaje natural.
- **tfidf**: Transforma una matriz de conteo a una representación tf o tf-idf [24] normalizada para recoger las caracteristicas del texto.
- **clf**: Inicializa el clasificador. [25]

Una vez creado el pipeline, el siguiente paso es aplicar cross-validation con un total de 10 veces, donde al final se realiza la media de las métricas: cobertura, precisión,medida F y exactitud descritas anteriormente. Dando como resultado:

SVM				
Promedio	Cobertura	Precisión	Medida F	Exactitud
MACRO	0.82	0.8383	0.8254	0.9485

Bayes

Es un clasificador basado en el teorema de Bayes, este teorema fue planteado por primera vez por Thomas Bayes en 1763. Es un teorema basado en la probabilidad condicionada. [26] de un evento A dado B. Este clasificador consiste en la asignación de un objeto que predice una serie de características o propiedades. De manera que predice si un objeto pertenece a una propiedad o a otras.

Para la creación de este clasificador, su desarrollo es similar a SVM con la diferencia que cuando se crea con un clasificador de tipo Bayes (MultinomialNB() [27])

[23] https://en.wikipedia.org/wiki/Bag-of-words_model
[24] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.htmll
[25] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
[26] https://es.wikipedia.org/wiki/Probabilidad_condicionada
[27] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

```
 text_clf = Pipeline([
        ('vect', CountVectorizer()),
        ('tfidf', TfidfTransformer()),
        ('clf', MultinomialNB()),
    ])
])
```

Tras esto se llamaría al conjunto de pruebas con cross validation, dando la media de 10 iteraciones:

Bayes				
Promedio	Cobertura	Precisión	Medida F	Exactitud
MACRO	0.6838	0.6581	0.6568	0.8956

SVM vs Bayes

Una vez obtenido el resultado de ambos clasificadores es hora de comparar los resultados.

Comparativa				
Clasificador	Precisión	Cobertura	Medida F	Promedio
SVM	0.82	0.8383	0.8254	0.9485
BAYES	0.6838	0.6581	0.6568	0.8956

A primera vista se puede observar que el clasificador SVM obtiene un mejor resultado para todas sus métricas:

- **Precisión**: se puede ver que el campo «precisión» es mayor en el SVM, es decir, este clasificador clasifica mejor las muestras positivas que el clasificador Bayes.
- **Cobertura**: se puede ver que el campo «cobertura» es mayor en el SVM, es decir, este clasificador clasifica mejor las muestras negativas que el clasificador Bayes.
- **Medida F**: se puede ver que el campo «medida f» es mayor en el SVM, es decir, este clasificador obtiene mejor proporción de las muestras negativas que positivas que Bayes.
- **Exactitud** : se puede ver que el campo «exactitud» es mayor en el SVM, es decir, este clasificador tiene un mayor porcentaje de acierto que Bayes.

Comparando ambos resultados se obtiene que el clasificador de tipo SVM obtiene un resultado mejor que el clasificador de tipo Bayes, por lo que se hará uso del modelo generado con el clasificador SVM.

6.3. Análisis de categorías

Una vez obtenido el clasificador ya se puede obtener el vocabulario de cada categoría para efectuar un análisis de cada frase con su categoría para recoger las características de cada una. Para realizar este análisis se crea de nuevo un archivo .csv que contendrá: fecha, lugar, frase y categoría. De esta manera se puede dividir fácilmente las categorías con sus frases, además se le añade la fecha y el lugar para hacer un estudio por cada zona.

En primer lugar para la creación de este archivo se recoge la fecha y la provincia de cada informe. Esta información se encuentra en la cabecera de cada informe como se ve en la Figura 16. El nombre de la provincia lo recojo como está en el informe y la fecha lo transformo a un formato: YYYY/MM/DD.

```
AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE ALACANT/ALICANTE

DÍA 30 DE NOVIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL SÁBADO 1
```

Figura 16: Cabecera de informe

Una vez obtenido la fecha y la zona se fragmenta las oraciones esta vez sin procesos de PLN y con el clasificador para asociar directamente la categoría con la frase. Un ejemplo de este fichero se puede ver en la Figura 17.

```
provincia,fecha,frase,clase
ALACANT/ALICANTE,2018-11-30,Temperaturas mínimas en ascenso y máximas en descenso en el interior .,temperatura
ALACANT/ALICANTE,2018-11-30,Viento del noroeste moderado .,viento
ALACANT/ALICANTE,2018-12-03,Poco nuboso con algunos intervalos de nubes altas .,nubes
ALACANT/ALICANTE,2018-12-03,Temperaturas en ascenso .,temperatura
ALACANT/ALICANTE,2018-12-03,Viento de componente oeste flojo tendiendo a variable por la tarde .,viento
ALACANT/ALICANTE,2018-11-25,Cielo con intervalos de nubes medias y altas .,nubes
ALACANT/ALICANTE,2018-11-25,No se descarta alguna precipitación débil .,precipitacion
ALACANT/ALICANTE,2018-11-25,Temperaturas con pocos cambios .,temperatura
ALACANT/ALICANTE,2018-11-25,Viento de componente oeste en general moderado .,viento
ALACANT/ALICANTE,2018-12-16,Tendiendo a poco nuboso o despejado por la tarde .,nubes
ALACANT/ALICANTE,2018-12-16,Temperaturas en ascenso .,temperatura
ALACANT/ALICANTE,2018-12-16,Viento del oeste flojo a moderado .,viento
ALACANT/ALICANTE,2019-01-11,Cielo despejado o poco nuboso .,nubes
ALACANT/ALICANTE,2019-01-11,Temperaturas mínimas en descenso y máximas con pocos cambios .,temperatura
```

Figura 17: Fragmento del fichero csv zona-fecha-frase-categoría

Tras conseguir esta relación es hora de realizar el análisis por categoría. Este análisis contendrá los siguientes apartados:

1. Número de veces que aparece una categoría por zona.
2. Longitud media de la frase de una categoría por zona.

3. Frecuencia de de sustantivos, adjetivos, verbos y adverbios.
4. Ampliación del vocabulario de entrada de datos , es decir, la relación de datos-texto del MTM de la sección 6.1.

En el paso 3 para la frecuencia de palabras se va a sacar el top 15-10 de las palabras que más se repitan. En esta frecuencia de palabras se hará el análisis por cada categoría, en otras palabras, los resultados y palabras que aparezcan en esta frecuencia de palabras marcará la estructura y las palabras de la formación de las futuras frases.

Además para tener una descripción gráfica y ver con un rápido vistazo las palabras más frecuentes se realizará una nube de palabras por cada categoría. Esta nube de palabras se ha desarrollado con la librería wordcloud [28] de pyhton. Para esta frecuencia de palabras se va a efectuar los procesos PLN aplicados anteriormente en el clasificador. Estos procesos son tokenización(separar palabra por palabra), postagging (etiquetación de sustantivo, adjetivo, verbo y adverbios), y por último lematización.

La lematización se emplea para sacar el lema de cada palabra y así evitar que cuente la misma palabra como palabras distintas.Por último para el paso 4 correspondiente a la ampliación del vocabulario se va a realizar una comparación entre el vocabulario definido en el MTM y el obtenido en esté análisis. En esta comparación se va emplear un proceso PLN denominado **similitud semántica**.

Esta similitud semántica consiste en averiguar cuanto de parecido tiene una palabra respecto a otra. Este proceso se ha llevado a cabo a través de la librería Spacy con la función «similarity» [29]. Esta función realiza esta similitud a través de vectores de palabra a través del algoritmo «word2vec» [30]. Este algoritmo saca una similitud a través de la función matemática coseno que cuanto más cercano del 1 son más parecidas las palabras y cuanto más cercano a 0 son menos parecidas. A la hora de ampliar el vocabulario se va a coger el vocabulario mayor o igual que 0.75. Este vocabulario añadido a la hora de formar la frase se escogerá aleatoriamente con una función random de forma que no se genere el mismo texto siempre.

A continuación se harán en análisis de cada categoría siguiendo estos pasos.

[28] https://www.datacamp.com/community/tutorials/wordcloud-python
[29] https://spacy.io/usage/vectors-similarity
[30] https://en.wikipedia.org/wiki/Word2vec

6.3.1. Nubosidad

Las frases de que consta la categoría nubosidad son un total de **187** frases de las 820 frases disponibles en total.

Precipitación		
Ciudad	**Longitud media**	**Cantidad**
TENERIFE	74,4	60
ALACANT/ALICANTE	55,15	43
LA RIOJA	98	39
MADRID	120	45
Total	86,88	187

Tabla 16: Longitud media y cantidad de nubosidad

Viendo la tabla se comprueba que esta temática esta siempre en el informe ya que los días totales por cada provincia son de 41 días y todas salvo en La Rioja concuerdan este dato. En cuanto a la longitud media de cada frase es de 86,88 caracteres.

Su vocabulario más frecuente se puede ver en la siguiente imagen:

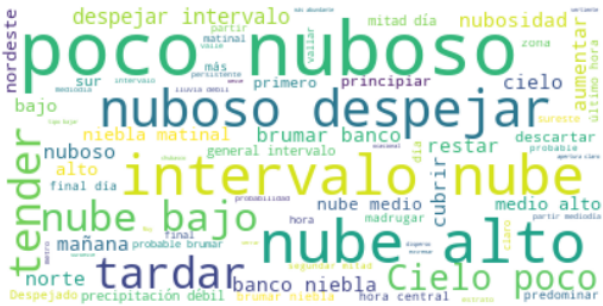

Figura 18: Nube de palabras para la categoría nubosidad

A continuación se va a analizar con más detalle la imagen indicando a la frecuencia de los sustantivos, adjetivos, adverbios y verbos:

Sustantivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
nube	113	día	25	norte	16
intervalo	80	hora	22	final	14
tarde	38	cielo	20	nuboso	11
niebla	30	resto	18	general	11
bruma	26	banco	18	mitad	9

Tabla 17: Frecuencia de sustantivos de la categoría nubosidad

Adjetivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
nuboso	167	medio	14	central	8
alto	91	débil	13	claro	6
bajo	43	probable	11	segundar	5
matinal	18	intervalo	9	persistente	3
cubrir	17	último	9	abundante	3

Tabla 18: Frecuencia de adjetivos de la categoría nubosidad

En los sustantivos aparecen periodos del día como «dia», «tarde» u «hora». Además de palabras con relación intima a nubosidad como «nube», «cielo» , «nuboso» o «claro». En cuanto a los adjetivos la mayoría son calificativos.

Adverbios			
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**
poco	104	también	2
más	10	encimar	2
principalmente	2	debajo	1
abriendose	2	muy	1

Tabla 19: Frecuencia de adverbios de la categoría nubosidad

Verbos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
tender	35	esperar	2	alternar	2
despejar	17	cubrir	2	romper	2
aumentar	16	crecer	2	persistir	2
descartar	10	incrementan	2	disminuir	1
predominar	7	cambiar	2	ocasionar	1

Tabla 20: Frecuencia de verbos de la categoría nubosidad

Respecto a los verbos se puede ver que a la hora de dar acción al sujeto, que en

este caso son las nubes, se puede dividir en tres tipos: subida, bajada y presentación del nubosidad.

Subida
aumentar
crecer
incrementan
cambiar
altenar

Tabla 21: Verbos que indican un aumento el grado de nubosidad

Bajada
despejar
disminuir
cambiar
altenar

Tabla 22: Verbos que indican una disminución del grado de nubosidad

Presentación
tender
predominar
esperar

Tabla 23: Verbos que indican presentación de nubosidad

Una vez visto las frecuencias de las palabras es hora de ver si se puede expandir el vocabulario de entrada, es decir, el de la relación datos-texto de la Sección 6.1.1. En esa sección el vocabulario del MTM es el siguiente vocabulario:

- Despejado
- Poco nuboso
- Intervalos nubosos/ Intervalos nubosos con lluvia/ Intervalos nubosos con nieve
- Nuboso / Nuboso con lluvia / Nuboso con nieve
- Muy nuboso / Muy nuboso con lluvia / Muy nuboso con nieve
- Cubierto / Cubierto con lluvia / Cubierto con nieve
- Nubes altas

Comparando con el vocabulario analizado en las distintas tablas de vocabulario analizadas en esta sección se puede ver que muchas de las palabras definidas en el MTM se encuentra como palabras de mayor frecuencia, por lo que es una confirmación que estas palabras deben de estar en los informes. En cuanto a palabras similares no hay ninguna palabra que se asemeje a este vocabulario por lo que no es posible expandir el vocabulario con la comparación de similitud semántica.

Conclusiones

Realizado el análisis es hora de sacar conclusiones. En primer lugar en la tabla de sustantivos aparece periodos del día por lo que se nombrarán junto al fenómeno meteorológico. Asimismo en los verbos aparecen verbos tanto de subida, de bajada y de presentación, así que a la hora de formar la frase se harán siguiendo este esquema: **fenómeno meteorológico + periodo del día + verbo** . En cuanto a la posibilidad de expandir el vocabulario de entrada no ha sido posible debido a que ninguna palabra se asemeja con el análisis descubierto.

6.3.2. Precipitación

Las frases incluidas en la temática precipitación son un total de **74** frases de las 820 frases disponibles en total. La cantidad de veces que aparece «precipitación» por zona y la longitud media se muestra en la siguiente tabla:

Precipitación		
Ciudad	**Longitud media**	**Cantidad**
TENERIFE	116,4	16
ALACANT/ALICANTE	93,1	7
LA RIOJA	121	30
MADRID	103	21
Total	104,6	74

Tabla 24: Longitud media y cantidad de la categoría precipitación

Según la tabla, la longitud media a la hora de realizar una frase es de 104,6 caracteres, además se puede apreciar que por la cantidad que aparece en los textos esta temática solo se informa cuando ocurre precipitaciones, siendo más frecuentes en climas subtropicales como Tenerife y climas fríos como Madrid y La Rioja y siendo la menor Alicante por ser un clima seco y cálido. Asimismo, el vocabulario más frecuente se resumen en esta imagen:

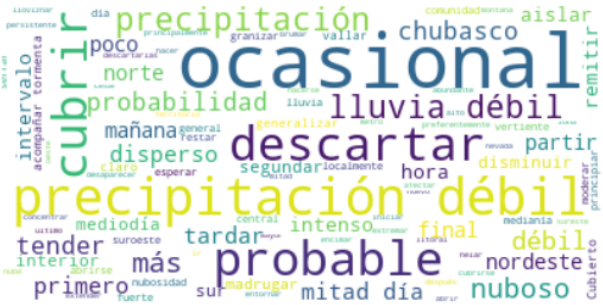

Figura 19: Nube de palabras para la categoría precipitación

Con mas detalle se puede ver en las siguiente tablas los sustantivos, adjetivos, adverbios y verbos más frecuentes:

Sustantivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
precipitación	26	tarde	6	nordeste	5
lluvia	13	final	6	intervalo	5
día	10	partir	5	norte	4
mitad	8	mañana	5	tormenta	4
probabilidad	10	hora	5	territorio	3

Tabla 25: Frecuencia de sustantivos de la categoría precipitación

Adjetivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
débil	33	primero	7	fuerte	3
ocasional	14	nuboso	6	acompañar	3
cubrir	13	disperso	6	flojo	4
probable	12	intenso	4	tormenta	4
chubasco	8	segundar	3	abundante	3

Tabla 26: Frecuencia de adjetivos de la categoría precipitación

En las tablas de sustantivos y adjetivos aparecen palabras relacionados con el periodo del día con palabras como «día», «mañana», «tarde» y «hora». Además y

lógicamente aparece sustantivos de precipitación como «precipitación», «lluvia» y «tormenta». En cuanto a los adjetivos se muestra adjetivos principalmente cuantitativos como «débil» o «fuerte».

Verbos			
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**
descartar	13	afectar	2
tender	8	desaparecer	2
remitir	4	aislar	2
disminuir	3	bajar	1
esperar	2	disipar	1

Tabla 27: Frecuencia de verbos de la categoría precipitación

En verbos se muestran principalmente verbos que indican una bajada de grado del fenómeno no mostrando ninguno de subida.

Adverbios			
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**
no	10	si	2
más	9	desaparecer	2
poco	5	preferentemente	1
localmente	2	después	1
principalmente	2	encimar	1

Tabla 28: Frecuencia de adverbios de la categoría precipitación

Una vez obtenida la frecuencia del vocabulario más usado es hora de aumentar el vocabulario de los datos de entrada. Este vocabulario sacado del MTM esta definido en la Sección 6.1.2. En la sección se habla de a la hora de describir la precipitación se hace a través : «lloviznas», «lluvias» , «chubascos» y «nieve». Además de hablar sobre «probabilidad-probabilidad nieve» ,«evolución temporal» y «distribución espacial».

Lloviznas, lluvias, chubascos y nieve

«Lloviznas», «lluvias», «chubascos» y «nieve» aparecen en las frecuencias de palabras de las tablas anteriores, por lo que también se confirma que el vocabulario definido en el MTM de la categoría precipitación es usado en los informes. Haciendo una comparación con los sustantivos de las tablas ninguno se asemeja a estos cuatros. Ahora bien a la hora de hablar de estos tres fenómenos se acompaña con adjetivos.

El vocabulario definido en el manual para describir estos acontecimientos es a través de adjetivos para dar significado al grado de la precipitación. Los adjetivos

son:

- «Débiles»
- Sin adjetivar (cuando es un valor intermedio)
- «Fuerte»
- «Muy fuerte»
- «Torrenciales»

En el análisis del texto de esta categoría se mostró los adjetivos más frecuentes. De estos adjetivos más frecuentes los que tienen más parecidos con la lista anterior es «intenso» y «flojo». Para saber si se puede añadir se va hacer el proceso PLN de similitud semántica. La comparación será «intenso» con «fuerte» y «torrenciales» ,y «flojo» con «débil».

Adjetivo 1	Adjetivo 2	Valor
débil	flojo	0.62

Tabla 29: Comparación del adjetivo «débil» de la categoría de precipitación

Adjetivo 1	Adjetivo 2	Valor
intenso	torrenciales	0.63
intenso	fuerte	0.83

Tabla 30: Comparación de adjetivo «intenso» de la categoría de precipitación

Dada la comparación de adjetivos el único que cumple un valor superior de 0.75 es «intenso» con «fuerte». Por lo que se añadirá «intenso» a la hora de hablar de precipitaciones fuertes.

Probabilidad

El vocabulario que especifica el MTM para describir la probabilidad de precipitación, de nieve y de tormentas es:

- «Baja probabilidad»
- «probable»
- «Sin adjetivar»

A simple vista no hay ningún adjetivo que sea parecido a la de la tabla de adjetivos número 26 . Por lo cual no se puede expandir este vocabulario.

Evolución temporal

El vocabulario por defecto usado para describir la evolución temporal es:

- «Ocasionales».
- «Persistentes» o «continuas»
- «Intermitentes».

Igualmente en la evolución temporal no hay ningún adjetivo que cuadre con los de la tabla de adjetivos .

Distribución espacial

El vocabulario por defecto usado para describir la distribución espacial es:

- «Aisladas» o «dispersas».
- «Sin adjetivar».
- «Generalizadas».

Igualmente que los dos anteriores, en los adjetivos de distribución espacial no hay ningún adjetivo que cuadre con los de la tabla de adjetivos.

Conclusiones

Realizado el análisis se va hacer una conclusión de la categoría de precipitación. En primer lugar en las tablas de sustantivos hay mención a los periodos del día por lo que se mencionaran junto a su fenómeno. En segundo lugar los verbos que más aparecen son de disminución de precipitación por lo que las frases irán en este sentido. Así que la estructura base de la frase será similar a la de la categoría nubosidad: **fenómeno meteorológico + periodo del día + verbo** En cuanto a la extensión de vocabulario de entrada de datos al final solo se añaden las palabras que se muestran a continuación:

Vocabulario por defecto	Vocabulario nuevo
intenso	fuerte

Tabla 31: Vocabulario ampliado de de entrada de datos de la categoría precipitación

6.3.3. Temperatura

Temperatura		
Ciudad	**Longitud**	**Cantidad**
TENERIFE	54,9	48
ALACANT/ALICANTE	53,4	57
LA RIOJA	60,7	79
MADRID	78,8	76
Total	61,95	260

Tabla 32: Longitud media y cantidad de la categoría temperatura

Las frases que se incluye en la temática d temperatura son un total de **260** frases de las 820 frases disponibles en total. Se puede decir que como supera los 41 días por zona aparece en todas los informes. Además su longitud media es de 61,95 siendo las frases con menor longitud de las frases para todas las categorías. Su vocabulario más frecuentes se resumen en la Figura 20.

Figura 20: Nube de palabras de la categoría temperatura

La imagen más detallada se puede encontrar en las tablas 33,34, 35 y 36 que muestran cuales son los sustantivos, adjetivos,verbos y adverbios.

Sustantivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
cambio	162	resto	34	heladas	12
Temperatura	160	interior	17	valle	12
descenso	113	zona	14	montaña	11
máximo	106	sierra	14	provincia	11
ascenso	91	generalizar	13	general	10

Tabla 33: Frecuencia de sustantivos de la categoría temperatura

Adjetivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
mínimo	98	acusar	10	notable	4
ligero	96	local	6	moderar	3
débil	64	significativo	6	bajo	3
importante	12	probable	5	especial	2
alto	12	intenso	4	grande	2

Tabla 34: Frecuencia de adjetivos de la categoría temperatura

En los sustantivos se muestra generalmente palabras asociadas a la temperatura como «temperatura» o «heladas» además de sustantivos donde se indican valores de ésta como «descenso», «ascenso» o «máximo». Los demás sustantivos están asociados a zonas como «interior», «sierra» o «montaña». Además se muestra que no hay ningún tipo de mención a los periodos del día. Los adjetivos se manifiestan mayoritariamente como calificativos como por ejemplo «débil», «grande» o «intenso».

Adverbios			
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**
más	13	principalmente	3
menos	5	no	2
localmente	4	solo	1
ligeramente	3	poco	1

Tabla 35: Frecuencia de adverbios de la categoría temperatura

Verbos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
alcanzar	3	dar	2	ascender	1
subir	3	experimentar	2	restar	1
mantener	3	descartar	1	helar	1
descender	3	disminuir	1	incrementar	1
predominar	2	aumentar	1	bajar	1

Tabla 36: Frecuencia de verbos de la categoría temperatura

En cuanto los verbos se presentan la mayoría con significados de subidas y bajadas de temperaturas además de presentación como «dar». Dando una distinción entre los verbos de subida y bajada:

Subida
subir
aumentar
ascender
incrementar

Tabla 37: Verbos que indican aumento de la temperaturaa

Bajada
descender
disminuir
restar
bajar
helar

Tabla 38: Verbos que indican una disminución de la temperatura

Una vez efectuado el análisis de los informes de temperatura se va hacer una comparación con el vocabulario de la relación datos-texto de la Sección 6.1.4. En dicha relación aparece dos tipos de vocabulario: uno de comparación de la temperatura frente al día anterior y otro para hablar de heladas.

Comparación de temperatura

- «Pocos cambios» o «cambios ligeros»
- «Aumento/descenso»
- «Aumento/descenso notable»
- «Aumento/descenso extraordinario»

Se puede ver que están los verbos «aumentar» y «disminuir». Dado tabla de verbos se va hacer una comparación semántica sobre estos.

Adjetivo 1	Adjetivo 2	Valor
aumentar	ascender	0,6
aumentar	subir	0m68
aumentar	incrementar	0,97

Tabla 39: Comparación del verbo «incrementar» en la categoría de temperatura

Adjetivo 1	Adjetivo 2	Valor
descender	disminuir	0,65
descender	bajar	0,78
descender	restar	0,58
descender	helar	0,44

Tabla 40: Comparación del verbo «descender» en la categoría de temperatura

Con el resultado de las tablas y aplicando el criterio de que se guarda solo los mayores de 0.75 se guarda para el verbo «aumentar» solo «incrementar» y para el verbo «descender» solo «bajar».

Una vez mirado los verbos vamos con los adjetivos. Se puede ver que en las viñetas numero 1, 3 y 4 están los adjetivos:«poco», «ligeros», «notables» y «extraordinario». Ahora bien mirando la lista de frecuencia de adjetivos los adjetivos «alto», «significativo», «intenso» y «grande» son los más parecidos a «notables» y «extraordinario». En cuanto al adjetivo «poco» y «ligero» tiene parecido con «débil» y «mínimo». Realizando la comparación con la métrica de similitud semántica se puede ver los siguientes resultados:

Adjetivo 1	Adjetivo 2	Valor
extraordinario	alto	0,72
extraordinario	intenso	0,59
extraordinario	significativo	0,66
extraordinario	grande	0,55

Tabla 41: Comparación del adjetivo «extraordinario» de la categoría temperatura

Adjetivo 1	Adjetivo 2	Valor
notable	alto	0,75
notable	intenso	0,67
notable	significativo	0,83
notable	grande	0,73

Tabla 42: Comparación del adjetivo «notable» de la categoría temperatura

En la tabla de «extraordinario» no hay ninguno que sea igual o mayor que 0.75 ,en cambio en la tabla «notable» si que cumple estos criterios los adjetivos «alto» y «significativo» por lo que estos dos se añaden a la hora de formar la frase. Ahora respecto a los adjetivos «poco» y «ligero»:

Adjetivo 1	Adjetivo 2	Valor
ligero	débil	0,66
ligero	mínimo	0,6

Tabla 43: Comparación del adjetivo «ligero» de la categoría temperatura

Adjetivo 1	Adjetivo 2	Valor
poco	débil	0,55
poco	mínimo	0,66

Tabla 44: Comparación del adjetivo «poco» de la categoría temperatura

Comprobando el resultado de los distintos adjetivos ninguno da un valor superior o igual al 0.75 por lo que no se ampliará ningún adjetivo en «ligero» o «poco».

Heladas

El vocabulario presentado por el MTM referente a heladas es:

- «débiles»
- Sin adjetivar (término medio)
- «Fuertes»

En primer lugar ya se ha hecho la comparación con «débil» la cual no daba ningún resultado favorable. En cuanto al adjetivo «fuerte»:

Adjetivo 1	Adjetivo 2	Valor
fuerte	intenso	0,83
fuerte	significativo	0,83
fuerte	grande	0,81
fuerte	alto	0.74

Tabla 45: Comparación del adjetivo «fuerte» en la categoría de temperatura

Viendo la comparación a la hora de hablar de heladas «fuertes» se podrá cambiar por heladas «intensas», «significativas» y «grandes».

Conclusiones

Hecho ya un análisis profundo del texto se puede sacar ya unas conclusiones. En primer lugar en las tablas de sustantivos no hay ninguna mención a los periodos del día, por lo tanto a la hora de hacer la frase no se harán menciones a ellos. En segundo lugar los verbos que más aparecen es para indicar aumentos o descenso de las temperaturas por lo tanto se enfocará la frase a este contexto. Además las frases tienen una longitud de media de 61,96 caracteres por lo que no puede ser muy extensa. En cuanto a la extensión de vocabulario de entrada de datos al final se queda de la siguiente manera:

Vocabulario por defecto	Vocabulario nuevo
aumentar	incrementar
descender	bajar
notable	alto, significativo
fuerte	intenso, significativo, grande

Tabla 46: Vocabulario ampliado de entrada de datos de la categoría temperatura

6.3.4. Viento

Las frases que se incluye en la temática de viento son un total de **254** frases de las 820 frases disponibles en total.

Viento		
Ciudad	**Longitud media**	**Cantidad**
TENERIFE	80,1	109
ALACANT/ALICANTE	56	48
LA RIOJA	66,25	48
MADRID	52	49
Total	63,56	254

Tabla 47: Longitud media y cantidad de la categoría viento

Viendo la tabla se comprueba que esta temática está siempre en el informe, como ocurre con las categorías «nubosidad» y «temperatura», ya que los días totales por cada provincia son de 41 días y en todas concuerdan este dato. En cuanto a la longitud media de cada frase es de 63.56 caracteres.

En cuanto a su vocabulario más frecuente:

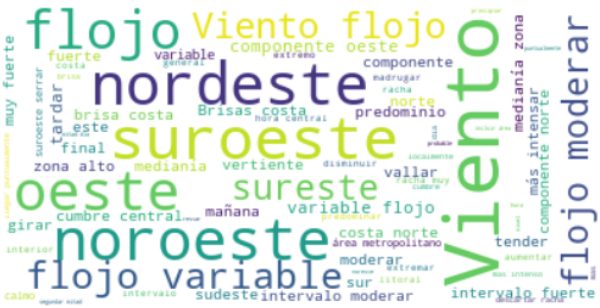

Figura 21: Nube de palabras de la categoría viento

Si desglosamos la imagen, podemos ver con más detalle las palabras utilizadas en las tablas 48, 49, 51 y 50 donde se indica la frecuencia de cada tipo palabra: sustantivos, adjetivos, verbos y adverbios.

Sustantivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
oeste	71	nordeste	44	medianía	20
vientos	59	intervalo	38	zona	19
tarde	52	sureste	30	cumbre	16
día	49	brisa	23	vertiente	15
componente	48	flojo	21	racha	15

Tabla 48: Frecuencia de sustantivos de la categoría viento

Adjetivos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
flojo	153	componente	17	incluir	3
moderado	63	central	16	orientar	3
variable	38	alto	15	ocasional	2
fuerte	38	metropolitano	8	debilitándose	1
bajo	18	intenso	4	persistente	1

Tabla 49: Frecuencia de adjetivos de la categoría viento

Los sustantivos muestran características del viento como su dirección con palabras como «oeste», «suroeste» o «nordeste», además de mencionar palabras realacionadas con ellas como «componente», «brisa», «racha» . Por último hay menciones a los periodos del día como «tarde» y «día». En cuanto a los adjetivos la mayoría muestra adjetivos calificativos.

Adverbios			
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**
más	25	no	3
muy	9	principalmente	2
puntualmente	7	intensificándose	1
localmente	4	asir	1

Tabla 50: Frecuencia de adverbios de la categoría viento

Verbos					
Palabra	**Frecuencia**	**Palabra**	**Frecuencia**	**Palabra**	**Frecuencia**
tender	20	llegar	4	dar	1
girar	11	descartar	4	excluir	1
predominar	6	incluir	2	cambiar	1
aumentar	5	arreciar	2	existir	1
disminuir	5	soplar	2	rolar	1

Tabla 51: Frecuencia de verbos de la categoría viento

Los verbos que se muestran se pueden dividir en subida, bajada, presentación del viento y asociados a la dirección:

Sube
aumentar
cambiar

Tabla 52: Verbos que indican un aumento de la categoría viento

Baja
arreciar
disminuir
cambiar
disminuir

Tabla 53: Verbos que indican disminución de la categoría viento

Igualdad
tender
predominar

Tabla 54: Verbos que indican presentación de la categoría viento)

Dirección
soplar
llegar
rolar

Tabla 55: Verbos de dirección de la categoría viento

Una vez descubierto qué vocabulario es más frecuente es hora de ver si se puede expandir con el vocabulario por defecto de la entrada de datos definido en la Sección 6.1.2. En esta sección se define dos tipos de vocabulario: uno para la dirección del viento, otro para la velocidad y rachas del viento.

Dirección del viento

En cuanto a la dirección del viento se encuentra el siguiente vocabulario:

- «Norte».
- «Noreste».
- «Este».
- «Sureste».
- «Sur».
- «Suroeste».
- «Oeste».
- «Noroeste».
- «Componente».

Comparando los sustantivos con los de la tabla, los sustantivos más parecidos es «componente» con «vertiente» y «medianía». Para ver si se añade este vocabulario se va a comparar con la técnica de similitud semántica.

Adjetivo 1	Adjetivo 2	Valor
componente	vertiente	0,62
componente	medianía	0,16

Tabla 56: Comparación del adjetivo «componente» de la categoría viento

Como se puede ver no se puede añadir ninguna palabra porque no son superiores al umbral del 0.75 definido en este trabajo.

Velocidad del viento

- «Calma».
- «Flojos».
- Sin adjetivad (términos medios)
- «Fuertes».
- «Muy fuertes».
- «Huracanados».

Puesto que las reglas definen un adjetivo para el viento según su grado se va a comprobar la tabla de adjetivos de la sección para ver si se puede ampliar este vocabulario. Los adjetivos más parecidos son: «moderado» e «intenso». Para el adjetivo «moderado» se podrá automáticamente en la viñeta número 3, ya que este se sitúa entre «flojos» y «fuertes». Ahora bien para el adjetivos «intenso» se va a aplicar el proceso PLN de similitud semántica entre «fuerte» y «huracanados»:

Adjetivo 1	Adjetivo 2	Valor
intenso	fuerte	0,83
intenso	huracanado	0,69

Tabla 57: Comparación de adjetivo «intenso» de la categoría viento

Conclusiones

En cuanto a las conclusiones, en primer lugar en la tabla de sustantivos aparecen palabras de periodos del día, así que se nombrará el fenómeno junto a su

periodo. En los verbos se muestra una bajada y subida de grado, además de verbos orientados a la dirección. Así que a la hora de formar la frase se hará a través de estos verbos para dar acción al sujeto, por lo que la estructura base de la frase será similar a los de la categoría nubosidad y precipitacón: **fenómeno meteorológico + periodo del día + verbo** . Por último las palabras añadidas al vocabulario de entrada son los siguientes.

Vocabulario por defecto	Vocabulario nuevo
intenso	fuerte
Sin adjetivar	moderado

Tabla 58: Vocabulario ampliado de entrada de datos de la categoría viento

6.3.5. Periodos del día

Por último aunque los periodos del día no es una categoría sí que está presente en todas las categorías excepto en la temperatura.A la hora de describir un fenómeno se tiene que nombrar el periodo del día que ocurre. En la Sección 6.1.7 se hizo una estimación de a qué rango de horas pertenece cada periodo del día. Ahora bien, puede ocurrir que ocurra el mismo fenómeno en todos los periodos del día, como por ejemplo que hayan cielos despejados todo el día, y puesto que las oraciones por el análisis realizado no deben de ser muy largas ya que la mayoría estan en torno a los 100 caracteres, se va a agrupar los periodos de está forma:

- **primera mitad del día**: cuando se repita un fenómeno en los periodos: madrugada,mañana y horas centrales del día. Esta agrupación se debe a que en la Tabla 48 aparece el sustantivo «mitad».
- **segunda mitad del día**: cuando se repita un fenómeno en los periodos: horas centrales del día,tarde y noche.
- **todo el día**: cuando un fenómeno ocurre en todos los periodos del día.
- **por el día**: cuando un fenómeno en los periodos: mañana,horas centrales del día y tarde. Esta agrupación se debe a que el MTM define día como la hora salida del sol y la hora de puesta de sol como se indica en la Sección 6.1.7.

6.4. Creación de reglas

La creación de reglas es la última fase para la generación del informe. Esta formación de reglas se ha realizado a través del estudio sobre el análisis del texto

de las categorías (Sección 6.3 y de la relación datos-textos (Sección 5.2).

La estructura que seguirán las frases serán: **Fenómeno meteorológico + periodo del día + verbo**. En los periodos del día se añadirá automáticamente para dar coherencia a la frase preposiciones que indica duración como: «por» o «durante». En cuanto a los verbos estarán todos en el tiempo verbal: **pretérito imperfecto de subjuntivo** debido a que este tiempo verbal puede expresar tanto pasado, presente y futuro, además de expresar hipótesis. Para presentar estos verbos se pondrá antes la conjunción **que**, por ejemplo: «que cesan», «que aumentan». Esta formación de preposiciones y conjunciones se explicará más al detalle más adelante.

Esta estructura estarán en todas las categorías salvo en «temperatura» ya que al realizar el análisis de vocabulario (Sección 6.3.3) no se mencionaba el periodo del día. Un ejemplo de un informe final creado automáticamente con datos de la comunidad de Madrid es el siguiente:

```
Cielos poco nubosos por la madrugada que incrementa a intervalos
nubosos durante la mañana y en las horas centrales del día
,y que disminuyen a cielos despejados por la tarde y por
la noche. Precipitaciones persistentes y generalizadas. Se
esperan lloviznas por la madrugada, lluvias por el día y
por la noche que remiten a lluvias débiles por la madrugada
,y chubascos por el día y por la noche. Temperaturas con cambios
ligeros con heladas débiles. Vientos que soplan del componente
este y sur que son, vientos flojos por la madrugada que arrecian
a vientos con calma por el día  y  que aumentan a vientos
flojos durante la noche.


TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):
Alcalá de Henares                 0     13
Aranjuez                         -1     14
Collado Villalba                  1     12
Getafe                           -1     13
Madrid                            1     13
Navalcarnero                     -2     13
```

Figura 22: Ejemplo de informe final

A continuación se explicarán las reglas que se han seguido para formar cada frase en cada categoría dando de ejemplo el conjunto de datos que forma el informe de la figura. Este conjunto de datos es obtenido de la base de datos.

6.4.1. Nubosidad

Recordando la Sección 6.1.1 la variable de entrada de está categoría se corresponde solo a la variable «estado del cielo» explicada en la Sección 5.

Por otro lado en el análisis, se ha descubierto que la categoría nubosidad es una de las categoría que aparece siempre en los informes como se ha visto en la Sección 6.3.1 con el número de veces que aparece. Las reglas de la formación de la frase de esta categoría se va a centrar en la información de la relación entre dato-texto de la Sección 6.1.1 y en el análisis del vocabulario de la Sección 6.3.1.

En la relación entre dato-texto se vio que la categoría nubosidad contiene datos ya en texto por lo que el dato concuerda con el texto. A la hora de forma la frase se va hacer siguiendo el análisis de la 6.3.1 que es nombrar el estado del cielo indicando su periodo de tiempo y nombrando sus cambios de grados. Para nombrar sus cambios de grado se harán a través de los verbos asociados a él que serán elegidos aleatoriamente.

Puesto que la frase tiene que tener una longitud cercana a 87 caracteres no se puede hacer muy extenso. Como el informe es a nivel de provincia que contiene generalmente varias zonas, es imposible hacer un informe detallado por lo que se va hacer es contar el periodo que más se repite en todas las zonas, es decir, el periodo general que ocurre en la provincia. Una vez obtenido esto la frase se hará nombrando la evolución de nubosidad pasando por las periodos del día, es decir, desde la madrugada hasta la noche. Los pasos llevados a cabo para la generación de la frase son los siguientes:

- Obtener de todas las zonas que forma la provincia el dato estado del cielo junto a sus horas. Un ejemplo de las 6 primeras horas de un día:

```
{'00': 'Despejado', '01': 'Despejado', '02': 'Despejado', '03': 'Despejado', '04': 'Intervalos nubosos', '05': 'Despejado', '06': 'Interval
{'00': 'Despejado', '01': 'Despejado', '02': 'Despejado', '03': 'Intervalos nubosos', '04': 'Despejado', '05': 'Poco nuboso', '06': 'Interv
{'00': 'Despejado', '01': 'Despejado', '02': 'Poco nuboso', '03': 'Despejado', '04': 'Despejado', '05': 'Despejado', '06': 'Intervalos nubo
{'00': 'Despejado', '01': 'Despejado', '02': 'Despejado', '03': 'Despejado', '04': 'Despejado', '05': 'Despejado', '06': 'Nuboso', '07': 'I
{'00': 'Poco nuboso', '01': 'Despejado', '02': 'Despejado', '03': 'Poco nuboso', '04': 'Nuboso', '05': 'Nuboso', '06': 'Intervalos nubosos'
```

Figura 23: Seis primeras horas de nubosidad

- Dividir los datos por periodos del día impuesto: madrugada, mañana, horas centrales del día, tarde y noche y obtener el fenómeno que más se repite junto a su periodo. En el caso que no haya algún fenómeno se etiquetará como False. Por ejemplo en este caso a no existir nubes con «lluvia» se etiqueta como False.

```
{'periodo': 'madrugada', 'general': 'Poco nuboso', 'lluvia': False, 'nieve': False}
{'periodo': 'mañana', 'general': 'Intervalos nuboso', 'lluvia': False, 'nieve': False}
{'periodo': 'horas centrales de la mañana', 'general': 'Intervalos nuboso', 'lluvia': False, 'nieve': False}
{'periodo': 'tarde', 'general': 'Despejado', 'lluvia': False, 'nieve': False}
{'periodo': 'noche', 'general': 'Despejado', 'lluvia': False, 'nieve': False}
```

Figura 24: Resumen de nubosidad con sus periodos de tiempo

- Obtener la evolución del fenómeno, es decir si baja, sube o si son iguales. En el caso que el fenómeno sea igual respecto al siguiente se juntarían con sus periodos de tiempo para no repetir información. Esto se puede ver en la comparación de las Figura 24 y 25 que muestran que durante las horas centrales del día y la mañana se repite el fenómeno de «Intervalos nubosos» y durante la tarde y noche el fenómeno «Despejado» .

```
['madrugada:Poco nuboso:sube', 'mañana,horas centrales de la mañana:Intervalos nuboso:baja', 'tarde,noche:Despejado:baja']
```

Figura 25: Agrupación de nubosidad con sus periodos de tiempo

- Formar la frase siguiendo la estructura: **fenómeno + fenómeno del día + verbo**. En la figura anterior (Figura 25) se puede ver que ya está los elementos básicos para la creación de esta estructura en los distintos elementos de la lista. El siguiente paso es transformar esta lista en las palabras que formaran la frases:

```
['cielos pocos nubosos:por la mañana:que crecen', 'intervalos nubosos:durante la mañana y en horas centrales del día: que se despejan', 'cielos despejados: durante la mañana y noche']
```

Figura 26: Ejemplo de construcción de frase de nubosidad

La formación de determinantes, sustantivos o conjunciones se ha formado de la siguiente manera:

- **Fenómeno**: para el fenómeno no se añade nada salvo la palabra «cielos» delante del adjetivo ya que el sujeto en este caso son los cielos. Por ejemplo si se da el caso de «fuertes» se añadirá cielos para completar el adjetivo como por ejemplo «cielos fuertes». En caso de que entre «intervalos nubosos» no se añade nada.
- **Periodo del día**: para el fenómeno del día se añade determinantes determinados y conjunciones. Para añadir el determinante determinado se tokeniza el periodo del día y se hace el proceso postagging. Si el periodo del día es de género femenino se añade «la» y si es de género masculino se añade «el». Una vez asociado el género se le añade aleatoriamente las preposiciones «por» o «durante» para dar coherencia de durabilidad.

Una vez añadido estas palabras a los periodos se guardan en una lista. De esta forma a iterar la lista se puede saber cual es el último elemento y añadir la conjunción «y» o en su defecto el signo de puntuación «,» para dar cohesión a este fragmento de la frase.

```
['durante la mañana', 'en las horas centrales del día']
durante la mañana y en las horas centrales del día
```

Figura 27: Ejemplo de formación de periodos del día

- **Verbos**: Por último los verbos están guardados en la forma de pretérito imperfecto de subjuntivo. Los verbos guardados son los analizados en el análisis de vocabulario de viento (Sección 6.3.1). En el caso que el verbo indica subida se añadirá un verbo de subida, si indica bajada de bajada y por último si indica presentación de presentación.

```
verbosIguales = [
    "que tienen","que predominan",
    "que esperan",
]

verbosDisminuir = [
    "que se despejan",
    "que disminuyen", "que cambian",
    "que alternan"

]

verbosAumentar = [
    "que crece","que aumenta",
    "que incrementa","que alternan",
    "que cambian"
]
```

Figura 28: Ejemplo de elección del verbos de nubosidad

- Por último queda unir la lista de la Figura 26 y dar coherencia en la frase se va a seguir el mismo proceso que en la temperatura. En el último elemento se añadirá la conjunción «,y» y a los demás la conjunción ¡a¿formando la siguiente frase:

```
Cielos poco nubosos por la madrugada que crece a intervalos nubosos durante la mañana y en las horas centrales del día ,y que se despejan a cielos despejados durante la tarde y por la noche
```

Figura 29: Ejemplo de frase nube

6.4.2. Precipitación

Recordando los datos de entrada que se vieron en la Sección 6.1.4 el único dato que forma la frase de temperatura se corresponde a «temperatura».

Recordando los datos de entrada de la categoría precipitación explicados en la Sección 6.1.2 son:

Categoría	Variables
Precipitación	precipitación
	probabilidad de precipitación
	nieve
	probabilidad de nieve
	probabilidad de tormenta

Tabla 59: Variables de entrada de la categoría precipitación

En cuanto a su análisis se averiguó que la categoría precipitación no está presente en todos los informes, sino que solo ocurre cuando hay precipitaciones tal y como se demostró en el análisis del texto 6.3.2, asi que cuando no ocurra este fenómeno no se formara la frase. Las reglas de la formación de la frase de esta categoría se van a centrar en la información de la relación entre dato-texto de la Sección 6.1.2 y en el análisis del vocabulario.

Siguiendo el esquema de la Sección 6.1.2, a la hora de formar la frase va a pasar por tres fases. Una primera fase donde se mire si existe precipitaciones, y si existe se clasifican en lloviznas, lluvias, chubasco o nieve tal y como se explica en la Sección 6.1.2. Una segunda fase donde se mire la probabilidad, distribución espacial y evolución temporal. Y por último, una última fase donde se mire si hay probabilidades de tormenta.

Para la formación de esta frase se va a seguir las conclusiones de su análisis de texto 6.3.2. En estas conclusiones se mostró que los verbos más frecuentes daban un significado de bajada de grado. Además de presencias de periodos del día. Así que para la formación de las frases se ordenará de mayor a menor el grado de la precipitación junto a sus periodos del día, para aplicar los verbos de bajada de precipitación. Por último la estructura de esta categoría seguirá siendo la básica de fenómeno + periodo del día + verbo con la excepción de indicar antes la probabilidad quedando de la siguiente manera: **probabilidad + fenómeno + probabilidad del día + verbo**.

- Obtener de la base de datos todas de las zonas el dato de precipitación, probabilidad de precipitación, probabilidad de tormenta , nieve y probabilidad de nieve junto a sus horas.

```
{'00': 'Ip', '01': '0.3', '02': '0.9', '03': '1.6', '04': '0.7', '05': '0', '06': '0.1',
{'00': 'Ip', '01': '0.8', '02': '0.1', '03': '0.2', '04': '0.5', '05': '0.8', '06': '0.8'
{'00': 'Ip', '01': '0', '02': '1.7', '03': '1.4', '04': '0.1', '05': '0', '06': '0.2', '0
{'00': '0.1', '01': '0.2', '02': '0.5', '03': '2.1', '04': '0.9', '05': '0.4', '06': '0.3
{'00': '0.3', '01': '1.4', '02': '1', '03': '0.3', '04': '0.1', '05': '0', '06': '0.4', '
{'00': 'Ip', '01': '0.6', '02': '0.7', '03': '0.1', '04': '0', '05': '0', '06': '0', '07'
```

Figura 30: Seis primeras horas de precipitación

```
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
```

Figura 31: Seis primeras horas de probabilidad de precipitación

```
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
```

Figura 32: Seis primeras horas de probabilidad de tormenta

```
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
```

Figura 33: Seis primeras horas de nieve

```
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
{'00': '0', '01': '0', '02': '0', '03': '0', '04': '0', '05': '0', '06': '0',
```

Figura 34: Seis primeras horas de probabilidad de nieve

- Se mira los datos de la probabilidad, distribución espacial, evolución tem-

poral ,probabilidad de tormenta y probabiliadd de nieve que según sus valores se adjetiva de una forma u otra.

```
Precipitaciones continuas y generalizadas.
```

Figura 35: Frase probabilidad

- Dividir los datos por periodos día: se agrupan los periodos del día en madrugada, mañana, horas centrales del día, tarde y noche y obtener el fenómeno que más se repite junto a su periodo. En el caso que no haya algún fenómeno se etiquetará como False.

```
{'periodo': 'madrugada', 'llovizna': 'lloviznas', 'generalLluvia': 'Débiles', 'generalChubascos': False, 'tormenta': False}
{'periodo': 'mañana', 'llovizna': False, 'generalLluvia': 'Lluvias', 'generalChubascos': 'Chubascos', 'tormenta': False}
{'periodo': 'horas centrales de la mañana', 'llovizna': False, 'generalLluvia': 'Lluvias', 'generalChubascos': 'Chubascos', 'tormenta': False}
{'periodo': 'tarde', 'llovizna': False, 'generalLluvia': 'Lluvias', 'generalChubascos': 'Chubascos', 'tormenta': False}
{'periodo': 'noche', 'llovizna': False, 'generalLluvia': 'Lluvias', 'generalChubascos': 'Chubascos', 'tormenta': False}
```

Figura 36: Resumen de precipitación con sus periodos de tiempo

- Ordenar los periodos de mayor a menor. Está ordenación es debido a que los verbos sacados de esta categoría describe un decremento del grado. Un ejemplo de esto ocurre en la siguiente figura donde se muestra ordenado «lluvia».

```
{'lloviznas': ['madrugada']}
{'Lluvias': ['mañana', 'horas centrales de la ', 'tarde', 'noche'], 'Débiles': ['madrugada']}
{'Chubascos': ['mañana', 'horas centrales de la ', 'tarde', 'noche']}
```

Figura 37: Lista ordenada de mayor a menor

- Formar la estructura general de: fenómeno + periodo del día + verbo. En esta ocasión no se indica en la lista bajada o subida de grado ya que esta frase se va a explicar en el sentido de disminución de precipitación. La estructura se ha formado de la siguiente manera:
  - **Fenómeno** : el fenómeno irá acompañado por tu tipo de precipitación, es decir, si es lloviznas, chubascos, lluvias o nieves. Además para nombrar el primer tipo de precipitación se usaran los verbos de presentación (Tabla 27) analizados en la Sección 6.3.2.

```
verbosPresentacion = [
    "se esperan","que afectan",
]
```

Figura 38: Verbos que presentan viento

De esta forma se le mas cohesión al fenómeno. Un ejemplo sería: «Se esperan chubascos...»

- **Periodo del día**: los periodos del día funciona de la misma manera que la categoría nubosidad, se añade preposiciones y determinantes para dar cohesión a la frase.
- **Verbo**: Por último los verbos, de igual manera funciona igual que la categoría nubosidad, se escogen los verbos analizados del vocabulario que en este caso coincidían solamente con una disminución del grado.

```
verbosDisminuir = [
    "que disminuyen","que disipan",
    "que remiten","que bajan",

]
```

Figura 39: Verbos de disminución del viento

- Por último formar la frase juntando la probabilidad + fenomeno + periodo del día + verbo. Para dar coherencia a la frases como ocurre en la categoría nubosidad se le añade las conjunciones «y» para dar paso al último elemento o «a» para dar paso al siguiente elemento.

```
Precipitaciones continuas y generalizadas. Se esperan lloviznas por la madrugada, lluvias por el día y durante la noche que remiten a lluvias débiles por la madrugada ,y chubascos por el día y por la noche.
```

Figura 40: Lista ordenada de mayor a menor sobre el grado de precipitaciones

6.4.3. Temperatura

La temperatura es una de las categoría que aparece siempre en los informes. Esta categoría tiene la característica de que no se le menciona los periodos del día, sino que trata de una evolución general de las temperaturas. De este modo las reglas de la formación de frase de esta categoría se va a centrar en la información dada en la Sección 6.1.4 que trata sobre:

- Cambios de temperatura respecto al día anterior
- Presencia de heladas.

Además de estos temas, se va añadir añadir un tema que se encuentra en todos los informes que es la tabla de temperaturas con las distintas zonas indicando el máximo y mínimo.

A la hora de formar la frase no se nombra ningún periodo del día debido a que en el análisis no se mencionan. Los pasos a seguir para la formación de la frase:

1. Obtener las máximas y mínimas de las zonas tanto del día de hoy como el día anterior.

```
['Alcalá de Henares:13:0', 'Aranjuez:14:-1', 'Collado Villalba:12:1', 'Getafe:13:-1', 'Madrid:13:1', 'Navalcarnero:13:-2']
['Alcalá de Henares:11:0', 'Aranjuez:11:0', 'Collado Villalba:13:2', 'Getafe:11:0', 'Madrid:11:2', 'Navalcarnero:12:-1']
```

Figura 41: Máximas temperaturas y mínimas temperaturas

2. Comprobar la diferencia entre las máximas y mínimas del día anterior. La estructura de está frase será: verbo + «con» + fenómeno. Los verbos serán elegidos aleatoriamente según el estudio de similitud semántica de la Sección 6.3.3.

```
if verbo == "sube":
    verbos = ["aumenta", "incrementa"]
    verbo = random.choice(verbos)
else:
    verbos = ["descenso", "bajadas"]
    verbo = random.choice(verbos)
```

Figura 42: Verbos de la categoría temperatura

```
bajadas de máximas
```

Figura 43: Ejemplo de frase de cambios de temperaturas

3. Comprobación de heladas. Estas heladas se comprueba con los mínimos de cada zona que compone la provincia. La estructura de la frase es: «con» + heladas.

```
heladas débiles
```

Figura 44: Ejemplo de frase de heladas de la categoría temperatura

4. La unión de las frases, en esta unión se empieza con la palabra «Temperatura» puesto que se habla de esta categoría.

```
Temperaturas con débiles cambios con heladas débiles.
```

Figura 45: Ejemplo de frase final de la categoría temperatura

5. Realizar la tabla de temperaturas por zona.

```
TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):
Alcalá de Henares                  0     13
Aranjuez                          -1     14
Collado Villalba                   1     12
Getafe                            -1     13
Madrid                             1     13
Navalcarnero                      -2     13
```

Figura 46: Tabla de temperaturas

6.4.4. Viento

Recordando las variables de viento de la Sección 6.1.5 que forman esta frase son:

Categoría	Variables
Viento	dirección del viento
	velocidad del viento
	racha máxima

Tabla 60: Variables de entrada de la categoría viento

Realizado el análisis se puede ver que la categoría viento es una de las categoría que aparece siempre en los informes .Las reglas de la formación de la frase de esta categoría se va a centrar en la información dada en la Sección 6.1.5 que habla de estas temáticas:

1. Dirección del viento.
2. Velocidad del viento.
3. Rachas máximas.

A la hora de formar la frase se hará a través del análisis de la categoría. Se mencionara el viento junto a sus periodos del día y se usarán los verbos asociados a la dirección del viento, subida y bajada e igualdad del viento. Por lo tanto la estructura será similar a las categorías nubosidad y precipitación: fenómeno + periodo del día + verbo con la diferencia que aquí primero se nombra la dirección del viento quedando la estructura de la siguiente manera: **dirección + fenómeno + periodo del día + verbo**. Los pasos a seguir para formar la frase son:

1. Obtener de todas las zonas el dato de la dirección, velocidad y rachas máximas junto a sus horas. Un ejemplo de las 6 primeras horas de un día:

```
{'00': 'S', '01': 'SO', '02': 'SO', '03': 'NE', '04': 'E', '05': 'NE', '06':
{'00': 'SE', '01': 'S', '02': 'S', '03': 'S', '04': 'SE', '05': 'SE', '06': '
{'00': 'SO', '01': 'SO', '02': 'O', '03': 'N', '04': 'E', '05': 'NE', '06': '
{'00': 'S', '01': 'O', '02': 'NO', '03': 'NO', '04': 'E', '05': 'NE', '06': '
{'00': 'SO', '01': 'SO', '02': 'SO', '03': 'S', '04': 'SE', '05': 'E', '06':
{'00': 'S', '01': 'S', '02': 'S', '03': 'S', '04': 'S', '05': 'S', '06': 'S',
```

Figura 47: Seis primeras horas de dirección del viento

```
{'00': '5', '01': '3', '02': '2', '03': '2', '04': '3', '05': '1', '06': '1', '
{'00': '7', '01': '6', '02': '4', '03': '2', '04': '3', '05': '4', '06': '5', '
{'00': '5', '01': '4', '02': '2', '03': '1', '04': '2', '05': '1', '06': '1', '
{'00': '1', '01': '1', '02': '5', '03': '4', '04': '1', '05': '2', '06': '1', '
{'00': '5', '01': '3', '02': '1', '03': '1', '04': '3', '05': '5', '06': '5', '
{'00': '4', '01': '4', '02': '4', '03': '3', '04': '3', '05': '4', '06': '6', '
```

Figura 48: Seis primeras horas de velocidad del viento

```
{'00': '8', '01': '5', '02': '3', '03': '4', '04': '5', '05': '1', '06':
{'00': '11', '01': '10', '02': '7', '03': '3', '04': '5', '05': '6', '06'
{'00': '8', '01': '7', '02': '3', '03': '1', '04': '4', '05': '2', '06':
{'00': '2', '01': '2', '02': '7', '03': '6', '04': '2', '05': '3', '06':
{'00': '8', '01': '5', '02': '3', '03': '2', '04': '4', '05': '8', '06':
{'00': '7', '01': '6', '02': '6', '03': '4', '04': '4', '05': '6', '06':
```

Figura 49: Seis primeras horas de racha máxima del viento

2. Transformar los datos en texto y dividir los datos por periodos: madrugada, mañana, horas centrales del día, tarde y noche y obtener el fenómeno más general junto a su periodo. El atributo «generalRacha» se encuentra en False si no supera el valor de 70 tal y como se menciona en la Sección 6.1.5. En caso contrario se le añade el adjetivo por defecto.

```
{'periodo': 'madrugada', 'generalVelocidad': 'Flojos', 'generalDireccion': 'S', 'generalRacha': False}
{'periodo': 'mañana', 'generalVelocidad': 'Calma', 'generalDireccion': 'SE', 'generalRacha': False}
{'periodo': 'horas centrales del día ', 'generalVelocidad': 'Calma', 'generalDireccion': 'NE', 'generalRacha': False}
{'periodo': 'tarde', 'generalVelocidad': 'Calma', 'generalDireccion': 'E', 'generalRacha': False}
{'periodo': 'noche', 'generalVelocidad': 'Flojos', 'generalDireccion': 'SE', 'generalRacha': False}
```

Figura 50: Resumen de la categoría viento con sus periodos de tiempo

3. Obtener la evolución del fenómeno, es decir si baja a o sube o si los periodos son iguales de forma similar a la variable nubosidad.

```
['madrugada:Flojos:baja', 'mañana,horas centrales del día ,tarde:Calma:sube', 'noche:Flojos:sube']
```

Figura 51: Agrupación de viento con sus periodos de tiempo

4. Obtener la estructura de la frase: dirección + fenómeno + periodo del día + verbo.
   - **Dirección**: para la dirección se mira las direcciones obtenidas y se transforma en su texto correspondiente tal y como se indican en la Sección 6.1.5. En el caso de que formen un componente se juntan las direcciones para describirlo. De la misma manera que se ha hecho anteriormente en el caso de que haya más de una dirección del viento se añade el signo de puntuación «,» y en el último elemento la conjunción «y».

```
['S', 'SE', 'NE', 'E', 'SE']
componente este y sur
```

Figura 52: Ejemplo de frase de dirección del viento.

   - **Fenómeno**: Para el fenómeno se le añade a todos los adjetivos el sustantivos «viento» puesto que se esta hablando de esta categoría. Por ejemplo si en los datos de entrada tenemos el dato 45 km/h que se corresponde al adjetivo «fuerte» se formaría con «vientos fuertes».
   - **Periodo del día**: El periodo del día se formaría de la misma manera que la categorías nubosidad y precipitación.
   - **verbo**: Por último el verbo se escogerá de los verbos analizados en la Sección 6.3.4

```
verbosDireccion = [
    "que llegan","que soplan",
    "que rolan",
]


verbosIguales = [
    "que tienden","que predominan",
]

verbosDisminuir = [
    "que disminuye",
    "que arrecian","que cambian"

]

verbosAumentar = [
    "que cambian"," que aumentan",
]
```

Figura 53: Verbos de la categoría viento.

5. Formación de frase con la siguiente estructura: «Vientos» + verbo de dirección viento + «que son» + fenómeno + periodo del día + verbo.

```
Vientos que rolan del componente este y sur que son, vientos flojos por la madrugada que arrecian a vientos con calma por la mañana  y  que aumentan a vientos flojos durante la noche
```

Figura 54: Ejemplo de frase de la categoría viento

7. Evaluación

Completado ya la creación de informes automáticos es hora de realizar comparaciones con los informes escrito por los profesionales. Para esta comparación se va a efectuar de nuevo la similitud semántica de la librería Spacy. Ya que esta metodología da como resultado si los textos tienen un significado similar.

Además se va hacer otro tipo de comparación. Se hará la comparación de categorías entre los distintos informes para ver si ambos hablan de la mismas categorías, además se va a comparar la longitud de frase por categoría. Esta comparación se hará gracias al clasificador.

Para realizar las comparaciones se va a elegir 8 tipos de informes con distintos fenómenos meteorológicos. La elección de estos 8 informes se va a dividir en 2 por cada provincia. Esta división será al azar y los mismos días para todas las provincias. De esta manera se puede obtener una comparación de las distintas zonas y ver si en algunas zonas se asemeja más o menos.

En estas comparaciones los informes realizados a mano por especialistas tendrá un nombre de «informe original», mientras que los automáticos: «informe automático».

7.1. Alicante

Los siguientes informes han sido generados para el día 1/12/2018

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE ALACANT/ALICANTE

DÍA 30 DE NOVIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL SÁBADO 1

ALACANT/ALICANTE

Cielo poco nuboso con intervalos de nubes altas. Temperaturas mínimas en descenso y máximas sin cambios. Viento de componente oeste flojo, con intervalos de moderado en zonas del interior.

Figura 55: Informe original del día 1/12/2018 de Alicante

Informe automático realizado para el día 2018-12-01
para la provincia de ALACANT/ALICANTE

Cielos poco nubosos durante todo el día. Temperaturas con débiles cambios. Vientos que rolan del oeste y noroeste que son, vientos flojos durante todo el día.

Figura 56: Informe generado automáticamente del día 1/12/2018 de Alicante

La semejanza con la comparación de similitud semántica es de 0.78 así que se asemejan bastante En cuanto a las categorías y sus longitudes se puede ver en las siguientes tablas:

Categoría	Longitud
nubosidad	41
temperatura	48
viento	78

Tabla 61: Categorías del informe original del día 1/12/2018 de Alicante

Categoría	Longitud
nubosidad	34
temperatura	30
viento	69

Tabla 62: Categorías de informe automático del día 1/12/2018 de Alicante

Comparando las tablas se informan de las mismas categorías y en el mismo orden, en cuanto al informe original las frases son más largas respecto al informe generado automáticamente. Además se muestra una pequeña variación en las tablas de temperaturas mostrando que son 1 o 2 grados distintos, por lo que los datos que dan ellos en el la api no concuerdan al 100 % con los del informe realizado de forma manual.

La siguiente comparación va ser del día 8/1/2019:

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE ALACANT/ALICANTE

DÍA 7 DE ENERO DE 2019 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL MARTES 8

ALACANT/ALICANTE

Cielo poco nuboso o despejado, salvo a primeras horas, cuando habrá intervalos nubosos sin descartar alguna llovizna ocasional. Temperaturas sin cambios importantes, o ascensos locales. Heladas débiles en puntos del interior. Viento variable tendiendo a oeste y noroeste flojo.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Alicante/Alacant	5	17
Orihuela	5	16
Dénia	5	17
Elche/Elx	5	17
Alcoy/Alcoi	2	14

Figura 57: Informe original del día 8/1/2019 de Alicante

Informe automático realizado para el día 2019-01-08
para la provincia de ALACANT/ALICANTE

Cielos nubosos durante la madrugada que alternan a cielos despejados por el día y durante la noche. Precipitaciones ocasionales y generalizadas. Se esperan lloviznas por la madrugada. Temperaturas con cambios ligeros. Vientos que soplan del componente oeste que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Alcoy/Alcoi	3	14
Alicante/Alacant	5	17
Dénia	5	16
Elche/Elx	6	17
Orihuela	6	16

Figura 58: Informe generado automáticamente del día 8/1/2019 de Alicante

La comparación de similitud semántica es de 0,8, así que de nuevo ambos informes se asemejan bastantes. En cuanto a si hablan de la mismas categorías se muestran en las Tablas 63 y 64.

Categoría	Longitud
nubosidad	75
precipitación	42
temperatura	51
temperatura	34
viento	44

Tabla 63: Categorías del informe origina del día 8/1/2019 de Alicante

Categoría	Longitud
nubosidad	81
precipitación	41
precipitación	33
temperatura	30
viento	70

Tabla 64: Categorías del informe automático del día 8/1/2019 de Alicante

En esta comparación se puede ver que las categorías son las mismas y en el mismo orden pero no con la misma frecuencia. En la tabla del informe 63 se aprecia que se menciona dos veces la categoría temperatura exponiendo heladas cuando en la tabla del informe original no se ve temperaturas menores de 0 grados que es cuando aparecen las heladas. Otra diferencia respecto a la Tabla 64 es que en este se habla dos veces de precipitación mencionando la probabilidad mientras que en el original no.

7.2. Tenerife

Se va a empezar mirando los informes generados para el día 1/12/2018:

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA ISLA DE TENERIFE

DÍA 30 DE NOVIEMBRE DE 2018 A LAS 13:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL SÁBADO 1

TENERIFE

Intervalos nubosos en el norte y este, predominando los cielos poco nubosos en horas centrales. En el resto, poco nuboso o despejado. Baja probabilidad de lloviznas en el nordeste, sobre todo a primeras y últimas horas. Temperaturas con pocos cambios o en ligero descenso. Viento del nordeste de flojo a moderado, más intenso en la costa sudeste durante la segunda mitad del día. En el Teide, viento flojo de componente norte girando a este al final. Brisas en la costa suroeste.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Santa Cruz de Tenerife 18 23

Figura 59: Informe original del día 1/12/2018 de Tenerife

Informe automático realizado para el día 2018-12-01
para la provincia de TENERIFE

Cielos nubosos por la madrugada que crece a cielos muy nubosos durante la mañana que disminuyen a cielos nubosos en las horas centrales del día que alternan a cielos muy nubosos por la tarde ,y que cambian a cielos poco nubosos durante la noche. Precipitaciones continuas y generalizadas. Se esperan lloviznas durante todo el día. Temperaturas con débiles cambios. Vientos que rolan del este que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):
Santa Cruz de Tenerife 18 23

Figura 60: Informe generado automaticamente del día 1/12/2018 de Tenerife

Esta comparación es la que tiene un valor de similitud peor con un valor de 0.71. En cuanto a la comparación entre las categorías y la longitudes de las frases:

Categoría	Longitud
nubosidad	81
nubosidad	31
precipitación	72
temperatura	45
viento	88
viento	58
viento	24

Tabla 65: Categorías del informe original del día 1/12/2018 de Tenerife

Categoría	Longitud
nubosidad	208
precipitación	42
precipitación	35
temperatura	30
viento	60

Tabla 66: Categorías del informe automático del día 1/12/2018 de Tenerife

En estas categorías ambas informan de los mismos temas, sin embargo hay diferencias como la longitud de nubosidad entre las dos siendo en el informe au-

tomático más enredado de explicar. En cuanto a las demás categorías son parecidas, sin embargo, en el fenómeno viento del informe original da más información sobre el Teide.

La siguiente comparación se corresponde al día 8/1/2019:

AGENCIA ESTATAL DE METEOROLOGIA

PREDICCIÓN PARA LA ISLA DE TENERIFE

DÍA 7 DE ENERO DE 2019 A LAS 13:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL MARTES 8

TENERIFE

Intervalos de nubes medias y altas. No descarta la presencia de calima débil en el sur durante la primera mitad del día. Temperaturas con pocos cambios o en ligero descenso de las mínimas. Viento de componente este flojo a moderado en costas y del sureste moderado en medianías y zonas altas. Brisas en costas noroeste. En altas cumbres centrales, componente sur flojo a moderado.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Santa Cruz de Tenerife 16 21

Figura 61: Informe origina del día 8/1/2019 de Tenerife

Informe automático realizado para el día 2019-01-08
para la provincia de TENERIFE

Cielos despejados por la madrugada que incrementa a cielos poco nubosos por el día y durante la noche. Temperaturas con cambios ligeros. Vientos que soplan del este y sureste que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):
Santa Cruz de Tenerife 16 21

Figura 62: Informe generado automáticamente del día 8/1/2019 de Tenerife

Esta similitud semántica entre ambos informe es de 0.75. En cuanto a las categorías y sus longitudes:

Categoría	Longitud
nubosidad	30
visibilidad	69
temperatura	57
viento	86
viento	23
viento	52

Tabla 67: Categorías de informe original del día 8/1/2019 de Tenerife

Categoría	Longitud
nubosidad	85
temperatura	30
viento	68

Tabla 68: Categorías de informe automático del día 8/1/2019 de Tenerife

La diferencia mas clara entre los dos informes es que en el informe original se muestra la categoría visibilidad que dado los datos que proporciona AEMET es imposible de realizar. En cuanto a las demás categorías que se muestran son las mismas, sin embargo, en la temática viento del informe original se muestra con más detalle esta categoría dando más detalles de zonas como costas y de montañas.

7.3. Madrid

En primer lugar se va a comparar los informes del día 1/12/2018:

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE MADRID

DÍA 30 DE NOVIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL SÁBADO 1

MADRID

Poco nuboso o despejado con intervalos de nubes medias y altas, más frecuentes por la tarde. No se descarta alguna bruma o banco de niebla en el valle del Tajo o de sus tributarios. Temperaturas mínimas en descenso, menos acusado en la Sierra. Temperaturas máximas en aumento en la Sierra y con pocos cambios en el resto. Heladas débiles en amplias zonas del centro y norte de la provincia, más acusadas en zonas altas de la Sierra. Vientos flojos variables tendiendo a suroeste.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Madrid	1	12
Alcalá de Henares	0	13
Getafe	0	13
Collado Villalba	1	14
Navalcarnero	0	12
Aranjuez	0	14

Figura 63: Informe original del informe original del día 1/12/2018 de Madrid

Informe automático realizado para el día 2018-12-01
para la provincia de MADRID

Cielos despejados durante la primera mitad del día que incrementa a cielos poco nubosos durante la tarde y durante la noche. Temperaturas con débiles cambios con heladas débiles. Vientos que rolan del oeste y suroeste que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Alcalá de Henares	-1	12
Aranjuez	1	13
Collado Villalba	1	12
Getafe	0	12
Madrid	1	11
Navalcarnero	0	11

Figura 64: Informe generado automáticamente del día 1/12/2018 de Madrid

La comparación de ambos es bastante positiva obteniendo un resultado de de 0.82. En cuanto a las categorías y longitudes:

Categoría	Longitud
nubosidad	77
visibilidad	71
temperatura	53
temperatura	64
temperatura	91
viento	41

Tabla 69: Categorías de informe original del día 1/12/2018 de Madrid

Categoría	Longitud
nubosidad	93
temperatura	47
viento	70

Tabla 70: Categorías de informe automático del día 1/12/2018 de Madrid

Como viene pasando en otras comparaciones en el informe original esta la categoría visibilidad mientras que en el automático no. En cuanto a las demás frases son muy parecidas excepto en temperatura que en el informe original da más información de esta categoría en zonas de sierra. Esta información con el conjunto de datos actual no se puede realizar.

Los siguientes informes han sido generados para el día 8/1/2019:

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE MADRID

DÍA 7 DE ENERO DE 2019 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL MARTES 8

MADRID

Predominio de cielos despejados aumentando la nubosidad alta durante la tarde. Temperaturas mínimas en descenso en el valle del Tajo y con pocos cambios en el resto. Temperaturas máximas en descenso, menos acusado en la Sierra. Heladas débiles generalizadas, menos probables en el oeste de la Sierra y más intensas en los valles del Lozoya, del Tajo y del Henares. Vientos flojos variables tendiendo a componente norte por la tarde.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Madrid	1	12
Alcalá de Henares	-5	14
Getafe	-1	13
Collado Villalba	1	16
Navalcarnero	-1	12
Aranjuez	-6	12

Figura 65: Informe original del día 8/1/2019 de Madrid

Informe automático realizado para el día 2019-01-08
para la provincia de MADRID

Cielos despejados durante la primera mitad del día y durante la tarde que aumenta a cielos poco nubosos por la noche. Temperaturas con cambios ligeros con heladas débiles. Vientos que llegan del este, noreste y noroeste que son, vientos flojos durante la madrugada que disminuye a vientos con calma por el día y durante la noche.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Alcalá de Henares	-3	14
Aranjuez	-4	12
Collado Villalba	0	14
Getafe	-2	12
Madrid	1	12
Navalcarnero	-2	12

Figura 66: Informe generado automáticamente del día 8/1/2019 de Madrid

La similitud entre ambos informes da un resultado de 0,88 siendo demasiados parejos. En cuanto a las categorías y sus longitudes:

Categoría	Longitud
nubosidad	68
temperatura	71
temperatura	53
temperatura	113
viento	58

Tabla 71: Categorías de informe original del día 8/1/2019 de Madrid

Categoría	Longitud
nubosidad	97
temperatura	47
viento	130

Tabla 72: Categorías de informe automático del día 8/1/2019de Madrid

Comparando las categorías los dos informas de las mimas categorías y en el mismo orden, sin embargo, en el informe original se muestra más información de la temperatura. En estas frases de temperatura dan mas información a zonas que no se disponen en el conjunto de datos como el rió Tajo o en los valles y en la sierras.

7.4. La Rioja

Como en las demás zonas la primera comparación se corresponde para el día 1/12/2018:

AGENCIA ESTATAL DE METEOROLOGÍA

PREDICCIÓN PARA LA PROVINCIA DE LA RIOJA

DÍA 30 DE NOVIEMBRE DE 2018 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL SÁBADO 1

LA RIOJA

Poco nuboso aumentando a nuboso, predominando la de tipo medio y alto. No se descartan brumas y nieblas matinales. Temperaturas mínimas sin cambios o en ligero descenso, y máximas en ascenso. Heladas en la sierra, débiles en general. Vientos del suroeste en la Sierra y flojos variables en el Valle.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Haro	3	14
Calahorra	2	12
Logroño	2	13

Figura 67: Informe original del día 1/12/2018 de La Rioja

Informe automático realizado para el día 2018-12-01
para la provincia de LA RIOJA

Cielos nubosos durante la madrugada que disminuyen a cielos poco nubosos durante la mañana y en las horas centrales del día que aumenta a cielos muy nubosos por la tarde ,y que cambian a cielos despejados por la noche. Precipitaciones continuas y generalizadas. Se esperan lloviznas durante todo el día. Temperaturas con débiles cambios. Vientos que soplan del oeste, noroeste y sureste que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Calahorra	2	11
Haro	3	14
Logroño	2	13

Figura 68: Informe generado automaticamente día 1/12/2018 de La Rioja

La similitud de estos dos informes da de un total de 0.69. En cuanto a las categorías y sus longitudes:

Categoría	Longitud
nubosidad	59
visibilidad	37
temperatura	65
temperatura	35
viento	54

Tabla 73: Categorías de informe original día 1/12/2018 de La Rioja

Categoría	Longitud
nubosidad	180
precipitación	39
precipitación	35
temperatura	30
viento	78

Tabla 74: Categorías de informe automático día 1/12/2018 de La Rioja

Está comparación es la más distintas de todas. En la primera tabla aparece la categoría visibilidad que dado los datos proporcionados por Aemet no es posible realizar la frase. Por otro lado en el informe automático menciona la categoría precipitación con lloviznas mientras que en el informe original no menciona nada de esto. Por último en el informe original menciona temperaturas en la sierra que dado que en el conjunto de entrenamiento no esta está información no se puede realizar la frase.

La otra comparación entre informes se da para el día 8/1/2019:

AGENCIA ESTATAL DE METEOROLOGIA

PREDICCIÓN PARA LA PROVINCIA DE LA RIOJA

DÍA 7 DE ENERO DE 2019 A LAS 14:01 HORA OFICIAL

PREDICCIÓN VALIDA PARA EL MARTES 8

LA RIOJA

Nuboso. Temperaturas mínimas sin cambios o en ligero ascenso y máximas en ligero ascenso. Heladas en la Sierra. Vientos del norte y noroeste, con rachas fuertes en la Rioja Baja.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Haro	2	9
Calahorra	3	10
Logroño	2	10

Figura 69: Informe original del día 8/1/2019 de La Rioja

Informe automático realizado para el día 2019-01-08
para la provincia de LA RIOJA

Cielos nubosos por la madrugada que disminuyen a cielos poco nubosos durante la mañana y en las horas centrales del día ,y que cambian a cielos nubosos durante la tarde y durante la noche. Temperaturas con cambios ligeros. Vientos que llegan del oeste y noroeste que son, vientos flojos durante todo el día.

TEMPERATURAS MÍNIMAS Y MÁXIMAS PREVISTAS (C):

Calahorra	3	9
Haro	2	9
Logroño	2	9

Figura 70: Informe generado automáticamente del día 8/1/2019 de La Rioja

La similitud de estos dos informes da un total de 0.71. En cuanto a las categorías y sus longitudes:

Categoría	Longitud
nubosidad	7
temperatura	69
temperatura	18
viento	55

Tabla 75: Categorías de informe origina del día 8/1/2019 de La Rioja

Categoría	Longitud
nubosidad	155
temperatura	30
viento	70

Tabla 76: Categorías de informe automático del día 8/1/2019 de La Rioja

Comparando las tablas se puede ver una gran diferencia entre la longitud de nubosidad. En la tabla del informe original nombra solo la palabra «nuboso» sin mencionar nada más, mientras que en el informe realizado automáticamente muestra con más en detalle la evolución de las nubes. En cuanto a las categorías son las mismas en ambas pero en el informe original hablan más sobre temperatura mencionando heladas en las sierra, cosa que con el conjunto de datos obtenidos no se puede saber.

7.5. Conclusiones

Una vez comparado varios informes se puede llegar a sacar conclusiones sobre el informe realizado automáticamente. Dando una comparación entre la media de los dos días analizados por zonas obtenemos los siguientes resultados:

Ciudad	Porcetaje de Acierto
Alicante	79 %
Tenerife	73 %
Madrid	85 %
La Rioja	70 %
Media	76.75 %

Tabla 77: Comparaciones de los resultados entre los informes

Mirando las tablas los peores porcentajes obtenido son de las provincias de La Rioja y Tenerife, mientras los mejores resultados son de Madrid y Alicante. Aun así da una media de acierto favorable de un 76.75 %.

Posibles mejoras para aumentar este porcentaje con los datos disponibles es dar una evolución mas general, ya que, muchas veces el informe automático pecaba de dar muchos detalles a la evolución de fenomenos como nubosidad y viento.

Otra forma de aumentar este porcentaje es de disponer de más datos meteorológicos, puesto que en muchos informes daban más detalles del tiempo meteorológico en costas, sierras o valles.

Por otro lado se ha visto que en los informes informaban sobre la categoría «visibilidad». Esta categoría, a día de hoy no se puede realizar una frase puesto que, en los datos proporcionados por Aemet no informa sobre si hay visibilidad o no. Una forma de solucionar esto es dar de entrada tanto los datos como el informe escrito y utilizar el clasificador en este informe para averiguar si hay visibilidad. Sin embargo este método se saldría de la idea original del proyecto que es obtener un resumen automático solamente con datos.

En conclusión se puede decir que el resultado final es favorable debido a que el informe automático da una información general de la provincia informando sobre las mismas categorías que los escritos por especialistas.

8. Conclusiones

En esta sección se hará unas conclusiones a nivel de todo el proyecto, se valorará mi experiencia, mis conocimientos adquiridos y futuro trabajo de esta aplicación

8.1. Conclusiones respecto al objetivo y al aprendizaje

Recordando los objetivos marcados desde un principio por este proyecto fueron los siguientes:

1. Aprender nociones básicas sobre el procesamiento del lenguaje natural y de las tecnologías del lenguaje humano, orientadas a la generación de información.
2. Analizar y decidir las tecnologías más apropiadas para desarrollar el trabajo
3. Aprender y utilizar herramientas y librerías específicas de Procesamiento de Lenguaje Natural, así como frameworks para diseñar arquitecturas (por ejemplo, NLTK o Spacy).
4. Analizar y desarrollar un método que permita generar informes meteorológicos.
5. Evaluar el método propuesto para determinar la calidad y adecuación del informe.

En esta lista desde mi criterio he superado todos los apartados. En los objetivos de aprendizaje he aprendido como funciona el PLN que cuando empecé el proyecto no tenía ninguna idea sobre esta tecnología, sin embargo, actualmente se puede decir que tengo los conocimientos básicos para realizar cualquier análisis de texto y recoger sus características

Además he aumentado mi formación respecto al *machine learning* y que es una tecnología que está en auge ya que he desarrollado un clasificador de texto cuando nunca lo había hecho.

Por último en cuanto al aprendizaje he aprendido las características del lenguaje de programación *Python* que hoy en día es un lenguaje demandado en el mercado laboral por su facilidad al análisis de datos.

8.2. Trabajo futuro

En cuanto al trabajo futuro se ha quedado cosas sin desarrollar, mi idea inicial era desarrollar una aplicación de escritorio o un sitio web donde en tiempo real recogiera los datos de AEMET y al instante creará el informe meteorológico explicando el porque de cada palabra. Sin embargo debido a la falta de tiempo no me ha sido posible desarrollarla

Por otra parte para perfeccionar el informe se podría realizar a todos los niveles. En la Sección 5.2 se definieron tres 3 tipos de informes: a nivel de provincia, comunidad y nacional. Uno de los trabajos futuros podría ser analizar los dos tipos de niveles restante para realizar informes de todo tipo.

Por último se podría hacer un análisis más profundo sobre el vocabulario, ya que el vocabulario recogido se corresponde a la estación de invierno. Se podría hacer un análisis de todas las estaciones para desarrollar un mejor vocabulario ya que seguramente las palabras a explicar el tiempo sean distintas.

9. Bibliografía y referencias

[Cheng Xiang Zhai Sean Massung]. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining (ACM Books)* 1ª ed.

[Trevor Hastie , Robert Tibshirani Jerome Friedman]. *The Elements of Statistical Learning Data Mining, Inference, and Prediction* 2ª ed.

[Yishu Miao Phil Blunso. *Language as a Latent Variable: Discrete Generative Models for Sentence Compression]*

[Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Mausam, Alan Ritter, Stefan Schoenmackers, Stephen Soderland, DanWeld, FeiWu, Congle Zhang] . *Machine Reading at the University of Washington*

[Matthew Richardson, Christopher J.C. Burges Erin Renshaw. *MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text]*

[Julio Castillo1 , Marina Cardenas1 , Adrián Curti1 , Osvaldo Casco1 Martin Navarro1 , Nicolás Hernandez1 , Melisa Velazco1.] *Desarrollo de Sistemas de Análisis de Text*

[José-Miguel Túñez-López , Carlos Toural-Bran Santiago Cacheiro-Requeijo]. *Automated-content generation using news-writing bots and algorithms: Perceptions and attitudes amongst Spain's journalists*

[Dörr, Konstantin]. *Mapping the field of Algorithmic Journalism*

[David Caswel Konstantin Dörr]. *Automated Journalism 2.0: Event-driven narratives*

[Marta Vicente, Cristina Barros, Fernando S. Peregrino, Francisco Agulló Elena Lloret]. *Automated Journalism 2.0: Event-driven narratives*