



ugr | Universidad
de Granada

TRABAJO FIN DE MASTER

MÁSTER UNIVERSITARIO OFICIAL EN CIENCIA DE DATOS E INGENIERÍA
DE COMPUTADORES

**Análisis de técnicas de visualización de opiniones y
desarrollo de una librería para la generación de
gráficos sobre valoraciones de usuarios**

Autor

Jonathan Espinosa López

Directores

Antonio Gabriel López Herrera
Jesús Alcalá Fernández



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

—
Granada, julio 1 de 2018

Análisis de técnicas de visualización de opiniones y desarrollo de una librería para la generación de gráficos sobre valoraciones de usuarios

Jonathan Espinosa López

Palabras clave: minería de opiniones, técnicas de visualización, análisis visual, extracción de tópicos, análisis de productos, análisis de sentimientos

Resumen

Muchos esfuerzos han sido realizados para aprovechar las bondades de las técnicas de análisis visual en la minería de opiniones. La complejidad y el continuo aumento del volumen de opiniones en Internet, dificultan la representación de dicha información. Los métodos desarrollados están constituidos por propiedades y características, que no han sido documentadas rigurosamente. Identificar y reconocer el alcance de las propuestas, brinda un marco contextual para que investigadores maximicen el margen de mejora de dichas técnicas. Este trabajo de fin de máster se centran en el análisis de las técnicas existentes, para comprender ventajas, limitaciones y características de cada representación. Se construye una librería de software, que incorpore todo el flujo de procesos necesarios, para representar y contrastar, algunas de las técnicas más relevantes en un caso de aplicación real. La librería desarrollada es implementada en el contexto de comercio electrónico, donde se realiza un análisis de productos y sus características, entorno a las consolas de video juegos más vendidas de las últimas dos décadas.

Analysis of opinion viewing techniques, and development of a library for the generation of graphics on user opinions

Jonathan, Espinosa López

Keywords: opinion mining, visualization techniques, visual analysis, topic extraction, product analysis, sentiment analysis

Abstract

Many efforts have been made to take advantage of the benefits of visual analysis techniques in the mining of opinions. The complexity and the continuous increase in the volume of opinions on the Internet, make it difficult to represent such information. The developed methods are constituted by properties and characteristics, which have not been rigorously documented. Identifying and recognizing the scope of the proposals, provides a contextual framework for researchers to maximize the margin for improvement of these techniques. This project focuses on the analysis of existing techniques, to understand advantages, limitations and characteristics of each representation. A software library is constructed to analyze sentiments, that incorporates all the necessary process flow, to represent and contrast, some of the most relevant techniques in case of real application. The developed library is implemented in the context of electronic commerce, where an analysis of products and their characteristics is performed, around the most sold video game consoles of the last two decades.

Yo, **Jonathan Espinosa López**, alumno del MÁSTER OFICIAL EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI Y4976445H, autorizo la ubicación de la siguiente copia de mi Trabajo de Fin de Máster en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Jonathan Espinosa López

Granada a 1 de julio de 2018

D. **Antonio Gabriel López Herrera**, Profesor del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada y miembro del Instituto Interuniversitario Andaluz DaSCI (Data Science and Computacional Intelligence).

D. **Jesús Alcalá Fernández**, Profesor del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada y miembro del Instituto Interuniversitario Andaluz DaSCI (Data Science and Computacional Intelligence).

Informan:

Que el presente trabajo, titulado *Análisis de técnicas de visualización de opiniones*, ha sido realizado bajo su supervisión por **Jonathan Espinosa López**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada al 1 de julio del 2018.

Los directores:

Antonio Gabriel López Herrera Jesús Alcalá Fernández

Agradecimientos

Agradezco a mis padres, por su apoyo y confianza. A mi esposa e hija por brindarme infinita motivación y a mis compañeros y profesores que hicieron de este, un esplendido proceso de aprendizaje.

Índice general

1. Introducción	23
1.1. Objetivos	24
1.2. Metodología	25
1.3. Planificación	26
1.4. Estructura de la obra	29
2. Estado del arte de las técnicas de visualización en minería de opiniones	31
2.1. Estructura de los datos de entrada	32
2.1.1. Fuente de datos	32
2.1.2. Formatos de la fuente de datos	33
2.1.3. Complejidad de las opiniones	34
2.2. Extracción de información de las opiniones	34
2.2.1. Extracción de tópicos	34
2.2.2. Extracción de características	35
2.3. Clasificación de opiniones	38
2.3.1. Ontologías	38
2.3.2. Lexicón	38
2.3.3. Aprendizaje automático	38
2.3.4. Polaridad de opiniones	38
2.4. Técnicas utilizadas en minería de opiniones	42
2.4.1. NPL: Para extracción de conocimientos y evaluación de la polaridad	44
2.4.2. Sentiment lexicon: para evaluación de la polaridad de las frases	46
2.4.3. Algoritmos de clasificación: Para extracción de frases de interés y evaluación de sentimientos	46
2.4.4. Algoritmos de agrupamiento: para el análisis de interacciones	47
2.4.5. LDA para la extracción de tópicos	48
2.5. Técnicas de visualización	48
2.5.1. Análisis visual temporal	51
2.5.2. Análisis visual orientado a la comparación de temáticas: .	54

2.5.3. Análisis visual orientado a la comparación de productos y sus características	55
2.5.4. Análisis visual orientado a la comparación de sentimientos	58
2.5.5. Análisis visual orientado a la comparación de usuarios e interacciones	63
2.5.6. Análisis visual con enfoque geoespacial	63
2.5.7. Análisis visual de aspectos demográficos	65
3. Análisis de técnicas de visualización	69
3.1. Principales funcionalidades:	69
3.1.1. Comparaciones:	70
3.1.2. Interacciones y relaciones:	70
3.1.3. Comportamiento histórico (temporal):	71
3.1.4. Descripción y caracterización de distribuciones (Resumen):	72
3.1.5. Descripción de parámetros demográficos:	72
3.1.6. Representación geográfica (espacial):	72
3.2. Capacidad de las técnicas	73
3.2.1. Tipo de datos que pueden procesar:	73
3.2.2. Cantidad de datos que pueden representar:	74
3.2.3. Número de dimensiones que pueden representar:	74
3.3. Características subjetivas	75
3.4. Utilidad de técnicas de análisis visual	76
3.5. Retos y oportunidades	78
4. Librería de visualización de opiniones	81
4.1. Librería en Python de análisis visual	81
4.1.1. Pre procesamiento de datos	82
4.1.2. Extracción de tópicos	86
4.1.3. Extracción de características	87
4.1.4. Técnicas de visualización	89
5. Caso de aplicación: Consolas de vídeo juegos	107
5.1. Fuente de datos	108
5.2. Selección de productos	110
5.3. Análisis de características y tópicos	112
5.3.1. Consolas de salón	114
5.4. Análisis de características	120
5.4.1. Consolas de salón	120
5.4.2. Consolas portable	122
6. Conclusiones y trabajo futuro	127

Índice de figuras

1.1. Tabla de tareas y fechas de inicio y fin del proyecto. En Azul los hitos que representan los principales entregables.	27
1.2. Diagrama de Gantt del proyecto. En Azul los hitos que representan los principales entregables.	28
2.1. Opinion Ring: Distribución del espacio de opiniones [19].	39
2.2. Diagrama de barras: porcentaje de sentimientos por país sobre T20 [54].	40
2.3. Diagrama de Pie: Proporción de sentimientos [24].	40
2.4. Ejemplo del mapeo gráfico: SentiVis[17].	41
2.5. Mapeo gráfico: SentiVis[17].	41
2.6. Mapa de relaciones [84].	42
2.7. Opinion Seer: Comparación visual de dos grupos de clientes A y B [91].	42
2.8. Enfoques de técnicas: minería de opiniones [63].	43
2.9. Arquitectura general de propagación entre componentes lingüísticos [17].	44
2.10. Representaciones gramaticales entre palabras en una frase [72]. . .	45
2.11. Estructura de adjetivos bipolares [28].	46
2.12. Pseudo código buscador de orientación de palabras [28].	47
2.13. Red social del agrupamiento sobre el tema Venezuela apaga RCTV [94].	47
2.14. Diagrama de LDA [45].	48
2.15. Evaluación de métodos de visualización parte 1 [70].	49
2.16. Evaluación de métodos de visualización parte 2 [70].	50
2.17. Evaluación de métodos de visualización parte 3 [70].	50
2.18. Número de publicaciones simuladas (línea) y datos reales (Barras) [84].	51
2.19. Vistas coordinadas representando interacciones de términos positivos y negativos [11].	52
2.20. Términos extraídos de opiniones positivas y negativas mensualmente [11].	52

2.21. Difusión de opiniones de una consola de video juegos en el periodo comprendido entre mayo 15 al 29, cuando una versión del dispositivo fue anunciada [15].	53
2.22. Exploración de la burbuja de un usuario en el hilo de opiniones [12].	54
2.23. Visualización del Opinion Ring [19].	55
2.24. Captura de pantalla de la interfaz de usuario del prototipo “User”. Las palabras representan a los agrupamientos encontrados y el color los sentimientos [23].	56
2.25. Visualización del Opinion Observer [43].	56
2.26. Mapa de relación competitiva [92].	57
2.27. Tree Map: Resumen de reporte de impresoras [56].	57
2.28. Sentiview: Atributo astrolabe [84].	58
2.29. Circular Correlation Map: Comentarios de todos los clientes [11].	59
2.30. Resultados de opiniones positivas y negativas [72].	59
2.31. Distribución de emociones en los premios Oscars en Tweeter respecto a la palabra Boyhood [9].	60
2.32. Spider Char: Valor de emociones para el hashtag: AAPHelpLine [40].	60
2.33. Diagrama de pie: Proporción de sentimientos positivos, negativos y neutros [20].	61
2.34. Diagrama de barras: Comparación de cantidades de sentimientos registrados en Tweets sobre tres productos [20].	61
2.35. Aggregation Chart: Distribución de documentos de interés seleccionados por el usuario [39].	62
2.36. Modelo SVM para revisiones positivas: Puntos azules representan opiniones positivas y puntos amarillos negativas. El área azul identifica los límites predichos por el modelo SVM para opiniones positivas [11].	62
2.37. (a) Diagrama de Euler, (b) Diagrama Spherule [69].	63
2.38. Red de MySpace en temas de política [94]: Cada nodo representa un individuo dentro de la red, las aristas muestran las relaciones de los nodos. Los individuos ubicados más cerca de la red contienen el mayor número de conexiones.	64
2.39. Key term Geo Map: (a) término Case Manager, (b) término Hawai [68].	65
2.40. Visualización de temas durante el temblor en Nepal: cada color en la figura hace referencia a alguno de los temas de tendencia, comentados en redes sociales durante el temblor; donate, tend, water, blood y medical, las concentraciones con mayor cantidad de puntos representan los temas de primera necesidad, en cada región durante la tragedia [68].	65
2.41. Frecuencia y distribución en el tiempo de las palabras de cada tema descubierto [68].	66
2.42. OpinionSeer: Múltiples vistas de opiniones de clientes [91].	66
2.43. OpinionSeer: Múltiples vistas de opiniones de clientes [91].	67

2.44. DemographicVis: Pasando el ratón sobre la interfaz en el segmento de mujer: 21-29 es posible ver los temas de interés del grupo [18]. . .	68
4.1. Esquema de clases de la librería de análisis visual desarrollada.	82
4.2. Diagrama detallado de clases de la librería de análisis visual desarrollada.	83
4.3. Diagrama pie: (a) Distribución 5 categorías, (b) Distribución 3 categorías.	93
4.4. Diagrama pie: (a) Proporción cantidad de opiniones, (b) Proporción promedio valoración.	94
4.5. Diagrama pie: (a) Proporción Tópicos y valoración producto 0700099867, (b) Proporción Tópicos y valoración producto B000006P0K.	94
4.6. Diagrama barras: Valoraciones de producto 0700099867 (a) 3 categorías (b) 5 categorías.	95
4.7. Diagrama barras: Análisis de tópicos (a) Promedio de valoraciones (b) Cantidad de opiniones.	96
4.8. Diagrama barras: Comparación de valoraciones y cantidad de opiniones de productos.	96
4.9. Diagrama barras: Comparación de valoraciones producto por tópico.	97
4.10. Diagrama polar: Comparación de valoraciones producto.	98
4.11. Diagrama de mapa de árbol: Resumen de valoraciones de producto.	99
4.12. Stack plot: Cantidad de opiniones por tópico de tres productos.	100
4.13. Series temporales: Valoración promedio de opiniones por año.	101
4.14. Series temporales: Cantidad de opiniones por año.	101
4.15. Series temporales: Valoración promedio de opiniones por año.	102
4.16. Series temporales: Cantidad de opiniones por año.	102
4.17. Parallel plot: Valoración de opiniones por año, a 16 años y b 5 años.	103
4.18. Mapa competitivo: P1 vs P2 por cada tópico, a horizontal, b vertical.	104
4.19. Observador de opiniones, análisis de tópicos de tres productos.	105
5.1. Diagrama de barras, comparación de categorías de valoración de clientes.	115
5.2. Stack plot, comparación productos Ninendo Wii con las generaciones anteriores de sus competidores, las consolas Xbox y Play Station 2.	115
5.3. Stack plot, comparación productos Ninendo Wii y sus competidores, las consolas Xbox 360 y Play Station 3.	116
5.4. Serie temporal productos Xbox 360, PS3 y Wii. a)Valoraciones promedio, b) Cantidad de opiniones.	117
5.5. Diagrama de Pie Consola Nintendo Wii.	117
5.6. Diagrama de Pie consola de Sony: PS3.	118
5.7. Diagrama de Pie consola de Microsoft: Xbox 360.	118

5.8. Stack plot, comparación productos Ninendo Wii con las generaciones anteriores de sus competidores, las consolas Xbox y Play Station 2	119
5.9. Serie temporal, comparación de tópicos productos a)Xbox y b) Play Station 2	120
5.10. Representación de tópicos con el Opinion observer para Xbox 360, PS3 y Wii.	121
5.11. Representación del histórico de tópicos con el Parallel plot sobre el total de opiniones de Xbox 360, PS3 y Wii.	121
5.12. Representación de valoración promedio de características con el Opinion observer para Xbox 360, PS3 y Wii.	122
5.13. Representación de valoración ponderada de características con el Opinion observer para Xbox 360, PS3 y Wii.	123
5.14. Representación de valoración promedio de características con el Opinion observer para DSI, Vita y PSP.	123
5.15. Representación de valoración ponderada de características con el Opinion observer para DSI, Vita y PSP.	124
5.16. Representación de valoración promedio de características con el Competitive Chart para DSI y PSP.	125

Índice de tablas

2.1. Áreas y métricas para la evaluación de técnicas de visualización	49
3.1. Enfoques de técnicas de visualización parte 1: según el objetivo de visualización	77
3.2. Enfoques de técnicas de visualización parte 2: según el objetivo de visualización	77
3.3. Enfoques de técnicas de visualización parte 1: según el tipo de técnicas utilizado	78
3.4. Enfoques de técnicas de visualización parte 2: según el tipo de técnicas utilizado	78

Capítulo 1

Introducción

El uso de plataformas digitales para la gestión y el intercambio de productos y servicios está en aumento [89], cada vez son más los usuarios que optan por utilizar Internet como medio para el intercambio de bienes y de consulta para tomar su decisión de compra[46]. Con el desarrollo de la comunidad 3.0 redes sociales, foros y blogs empiezan a tomar el papel de agentes clave en la gestión de información. Internet se perfila como una plataforma de libre intercambio de opiniones, donde los usuarios describen con plena libertad sus críticas y comentarios acerca de productos y servicios, despertando el interés de analistas que quieren comprender este fenómeno, para encontrar la mejor manera de recuperar esta información, evaluarla y visualizarla.

El análisis de las opiniones de los clientes tiene un enorme campo de aplicación, donde se benefician por un lado los clientes, pudiendo establecer criterios que apoye en su decisión de compra, y por otro lado las empresas, donde estas obtienen un beneficio con el feedback de sus clientes, que favorecerá en la toma de decisiones dentro de la compañía. El resultado, una mejora en procesos tales como: el diseño y configuración de productos o servicios, marketing y publicidad, vigilancia tecnológica, aseguramiento de la calidad, seguimiento de la satisfacción de cliente y experiencia de usuario como en [8].

Las aplicaciones en minería de opiniones son extensas, aplicaciones en sitios web para sistemas de recomendación, inteligencia de negocios, marketing entre otras. El abanico de posibilidades es enorme en un mercado digital que cada vez es mas extenso [57]. La cantidad y diversidad de información de la que se dispone, plantea serios retos que nacen de la misma complejidad del ser humano y su lenguaje, planteando a los analistas un importante desafío en el campo visualización y la minería de opiniones.

Las técnicas de visualización adquieren protagonismo, gracias a su capacidad para transformar datos brutos en información, que a su vez se transforma en conocimiento. Los gráficos tienen una alta capacidad de simplificar

la información y permiten que esta se transfiera de forma muy intuitiva. En contextos de alta complejidad de datos, como en el análisis de opiniones, el problema de extracción de conocimiento es mucho mayor, debido a la estructura de las opiniones, que contienen un alto grado de espontaneidad, y el gran volumen de datos que tienen que procesar las compañías de comercio electrónico. En estas condiciones, se vuelve difícil la transferencia de información “compleja” a un usuario humano, incapaz de abstraer conocimiento de un gran volumen de opiniones. Las técnicas de visualización se presentan como una atractiva solución, que sirve de interfaz, entre espacios de datos de gran volumen y alta complejidad, con la extracción de conocimiento de opiniones.

A la fecha los métodos de análisis visual para la minería de opiniones no han sido documentadas rigurosamente. Esta tarea es clave, por un lado en el proceso de investigación, donde se hace necesario identificar y reconocer el alcance, impacto, restricciones, ventajas y características de las propuestas existentes. Por otro lado, este trabajo tiene relevancia en el campo de aplicación, donde expertos pueden aprovechar las bondades de cada técnica, para aportar el máximo valor a sus análisis.

Este trabajo de fin de Máster se centran en el análisis de las técnicas existentes, para comprender ventajas, limitaciones y características de cada representación. El objetivo final es construir una librería de software de libre distribución, que incorpore todo el flujo de procesos necesarios, para representar y contrastar, algunas de las técnicas más relevantes en un caso de aplicación real, con un conjunto de datos de la compañía estadounidense de comercio electrónico Amazon¹.

1.1. Objetivos

Objetivo general

Realizar una revisión bibliográfica de las distintas técnicas de análisis visual utilizadas en la literatura para realizar minería de opiniones. Analizar las ventajas y carencias de cada una de estas técnicas, y desarrollar una plataforma software de libre distribución que integre algunas de las técnicas más relevantes de la literatura. Esta plataforma permitirá analizar la información que proporcionan estas técnicas para realizar un análisis completo de opiniones.

¹Amazon, S.L. es una compañía estadounidense de comercio electrónico y servicios de computación en la nube a todos los niveles con sede en la ciudad estadounidense de Seattle.

Objetivos específicos

- Realizar un análisis bibliográfico de las distintas propuestas existentes en la literatura.
- Analizar las ventajas y carencias de las distintas propuestas
- Desarrollar una herramienta software de libre distribución que integre algunas de las propuestas más interesantes de la literatura.
- Realizar un análisis de un caso práctico, mediante las técnicas integradas en la herramienta software desarrollada.

1.2. Metodología

La metodología de este proyecto está definida en 9 etapas, donde las 4 primeras constituyen procesos de análisis y las últimas dos tratan aspectos de desarrollo se software con un caso de aplicación.

Las principales etapas se describen a continuación:

- Análisis bibliográfico: Esta etapa constituye una búsqueda de documentación en los temas en que convergen en analistas visual y la minería de opiniones. La documentación recolectada es filtrada y clasificada. Cada grupo de documentos es analizado contrastando las técnicas según su topología.
- Análisis de la técnicas de visualización de opiniones: En esta etapa el análisis bibliográfico se extiende en un plano más crítico, donde se estudia el alcance, ventajas, restricciones y oportunidades de las técnicas encontradas.
- Análisis de paquetes de software de libre acceso: De cara al desarrollo de la librería de análisis visual de este proyecto, se procede a realizar una búsqueda de paquetes de software de trabajos existentes y librerías de libre acceso, tanto para visualización, como pre procesamiento de datos, extracción de tópicos y características.
- Análisis y selección del conjunto de datos del caso de aplicación: Este proceso es responsable de la adquisición de un conjunto de datos que incluya opiniones y valoraciones de usuarios de productos ó servicios.
- Desarrollo de funciones de procesamiento de datos: Conociendo los requisitos de entrada de las librerías de visualización disponibles, y las estructuras de datos típicos de "datasets" de opiniones de productos y servicios, se desarrollan los métodos de transformación necesarios, que permitan implementar las librerías de visualización encontradas.

- Implementación de técnicas de extracción de información: Para poder dar un contexto a las técnicas de análisis visual de opiniones, es necesario definir un conjunto de características y tópicos sobre los cuales centrar el análisis. Para llevar a cabo esta tarea se implementará una técnica para extracción de tópicos y una de características, teniendo en cuenta lo visto en el análisis bibliográfico.
- Desarrollo de métodos para la construcción de gráficos: Sobre la base de librerías de visualización encontradas se construirán las funciones que formarán dichas estructuras para representar algunas de las técnicas de visualización más relevantes.
- Construcción de librería de análisis visual: En esta etapa los métodos desarrollados anteriormente se consolidan en un paquete con programación orientada a objetos.
- Caso de aplicación: Se implementa la librería de análisis visual en un conjunto de datos real, que fue seleccionado en la etapa de “análisis y selección del conjunto de datos del caso de aplicación”. Sobre el conjunto de datos seleccionado, se planteará un tema de análisis para contextualizar el caso de aplicación. Finalmente se describirán aspectos de interés encontrados.

1.3. Planificación

El alcance definido de este trabajo se enfoca principalmente en 3 entregables: un documento con el análisis de las técnicas existentes, una librería de análisis visual de opiniones de libre distribución, y una implementación de la librería en un caso de aplicación real. Cada entregable requiere del desarrollo de una serie de actividades, que se relacionan como se muestra en el diagrama de Gantt de la Figura 1.2. El tiempo propuesto de desarrollo es aproximadamente cinco meses (20 semanas). Los recursos técnicos disponibles comprenden un ordenador tipo laptop de referencia Dell Vostro 3550, el recurso humano está constituido por tres personas, dos profesores vinculados al proyecto en calidad de tutores, con una disponibilidad de 2 horas semanales, y un ingeniero investigador con una disponibilidad de 20 horas semanales. Los costos de software asociados son nulos, debido a que se utilizarán solo herramientas de libre distribución.

La distribución de los recursos disponibles y representación de los hitos que materializan los entregables se pueden observar en el diagrama de Gantt de la Figura 1.2, resaltados en azul. Los tiempos de inicio, fin y duración de cada actividad se muestran en la Figura 1.1, donde se define la fecha de inicio oficial el 1 de septiembre de 2017, y la fecha de fin del proyecto el 18 de enero de 2018.

WBS	Nombre	Inicio	Fin	Trabajo	Duración
1	▼ Análisis bibliográfico	sep 1	sep 21	26d	15d
1.1	Recolección de fuentes bibliográficas	sep 1	sep 1	1d	1d
1.2	Clasificación de fuentes bibliográficas	sep 1	sep 21	15d	15d
1.3	Lectura agil de fuentes encontradas	sep 1	sep 7	5d	5d
1.4	Filtrado de artículos más relevantes	sep 1	sep 4	5d	1d 5h
2	▼ Análisis de las técnicas de visualización de los datos	sep 22	oct 12	65d	15d
2.1	Lectura crítica de artículos filtrados	sep 22	oct 12	15d	15d
2.2	Documentar estructuras de fuentes de datos	sep 22	sep 28	5d	5d
2.3	Documentar técnicas de procesamiento de datos	sep 22	oct 12	15d	15d
2.4	Documentar topologías de las técnicas de visualización	sep 22	oct 12	15d	15d
2.5	Redactar un análisis crítico de las técnicas de visualización	sep 22	sep 28	15d	5d
3	▼ Análisis de paquetes de software de libre distribución	oct 13	nov 2	25d	15d
3.1	Identificar las librerías de libre distribución	oct 13	oct 26	10d	10d
3.2	Leer documentación asociada a las librerías	oct 13	nov 2	15d	15d
4	▼ Análisis y selección del conjunto de datos	sep 22	oct 19	30d	20d
4.1	Búscar fuentes de datos con opiniones de expertos	sep 22	oct 19	20d	20d
4.2	Analizar fuente de datos	sep 22	oct 5	10d	10d
5	Análisis de las técnicas existentes	nov 2	nov 2	N/D	N/D
6	▼ Implementación de técnicas de extracción	nov 3	nov 23	30d	15d
6.1	Implementar técnicas de extracción de tópicos	nov 3	nov 23	15d	15d
6.2	Implementar técnicas de extracción de caras	nov 3	nov 23	15d	15d
7	▼ Desarrollo de librería de análisis visual	nov 24	ene 4	70d	30d
7.1	Desarrollo de funciones de procesamiento	nov 24	ene 4	30d	30d
7.2	Desarrollo de métodos para la construcción	nov 24	ene 4	30d	30d
7.3	Estructurar librería en clases	nov 24	dic 7	10d	10d
8	Libería análisis visual	ene 4	ene 4	N/D	N/D
9	▼ Caso de aplicación	ene 5	ene 18	25d	10d
9.1	Definir contexto del análisis a realizar	ene 5	ene 18	10d	10d
9.2	Ánalisis y filtrado de datos	ene 5	ene 11	5d	5d
9.3	Implementar librería de visualización deseada	ene 5	ene 11	5d	5d
9.4	Analizar visualizaciones obtenidas	ene 5	ene 8	5d	1d 5h
10	Aplicación caso real	ene 18	ene 18	N/D	N/D

Figura 1.1: Tabla de tareas y fechas de inicio y fin del proyecto. En Azul los hitos que representan los principales entregables.

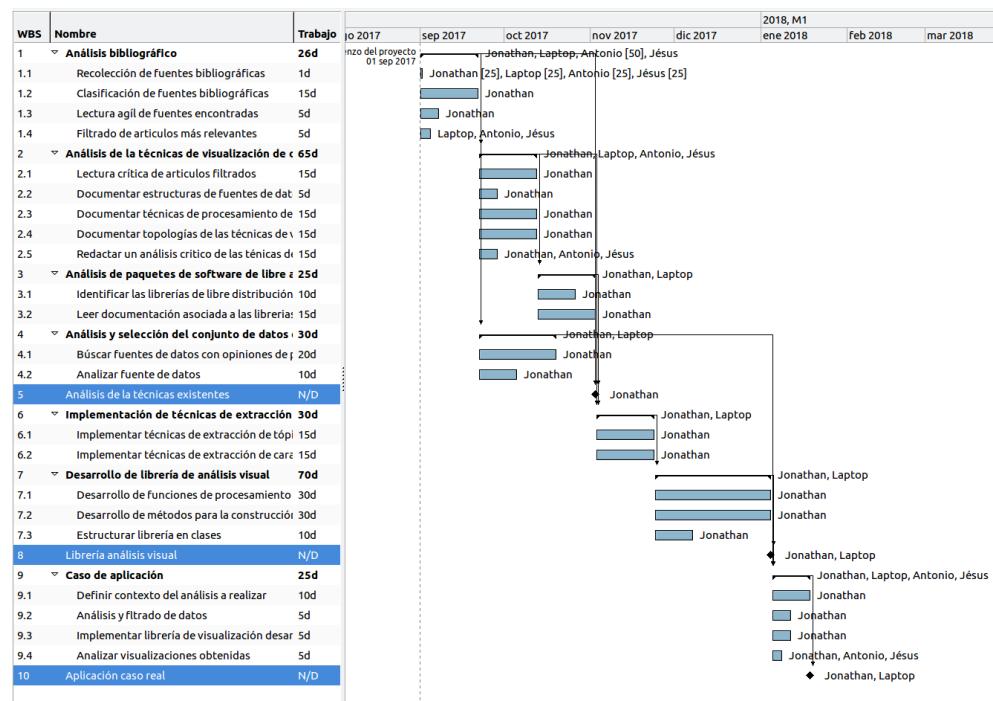


Figura 1.2: Diagrama de Gantt del proyecto. En Azul los hitos que representan los principales entregables.

1.4. Estructura de la obra

La obra se distribuye en 6 capítulos, que describen el desarrollo del planteamiento expuesto en la sección anterior. La descripción del contenido se describe a continuación:

- Capítulo 1 - Introducción: Contiene una descripción introductoria del problema a resolver y la estrategia de solución planteada.
- Capítulo 2 - Estado del arte: Describe los avances y novedades enmarcados en el análisis visual de opiniones. La descripción es realizada con la información abstraída de los artículos más destacables de la bibliografía. Contiene múltiples secciones donde se desglosa cada uno de los componentes en el flujo de datos necesario para implementar las técnicas de estudio.
- Capítulo 3 - Análisis de técnicas: Despues de recopilar los trabajos más relevantes en el contexto de visualización de opiniones, se procede a realizar un análisis critico de los trabajos de la literatura. En este capitulo se propone una topología de clasificación de las técnicas en una serie de áreas y dimensiones.
- Capítulo 4 - Librería análisis visual de opiniones: En este capítulo se describen los componentes principales de la librería desarrollada, ilustrando brevemente un ejemplo de aplicación de cada análisis visual implementado. Adicionalmente se al paquete de software se incluye la descripción de la implementación de extracción de características y tópicos desarrollada.
- Capítulo 5 - Caso de aplicación: Se plantea el caso de aplicación para analizar las consolas de video juegos más influyentes en el conjunto de datos de estudio. El capitulo desarrolla una implementación de la librería propuesta y un análisis de los resultados conseguidos, sobre algunas de las técnicas del paquete de software desarrollado.
- Capítulo 6 - Conclusiones: Se destacan los descubrimientos más relevantes, encontrados en el desarrollo de este proyecto.

Capítulo 2

Estado del arte de las técnicas de visualización en minería de opiniones

Se realizó una búsqueda bibliográfica, para identificar las principales técnicas de visualización empleadas en el ámbito de la minería de opiniones para el análisis de comentarios de usuarios. Las técnicas de visualización encontradas describen varias topologías, donde se identificaron varias etapas claves previas a la misma visualización, iniciando desde la recopilación de opiniones respecto a una temática particular, la extracción e identificación de las características a evaluar, la evaluación de dichas características, hasta los métodos aplicados para resumir y presentar la información en un gráfico.

Después de un análisis exhaustivo, dentro del marco de referencia se encontraron varias vertientes. Cada tipo de visualización pretende hacer un enfoque en un campo particular. Cada campo hace referencia a una dimensión que describe uno o varios conjuntos de atributos que muestran y comparan información acerca del usuario, del producto o servicio, la empresa, el mercado y la experiencia de usuario. Los datos empleados en el contexto de estudio provienen principalmente de cuatro fuentes; sitios web de comercio electrónico, redes sociales y en una menor medida de foros y blogs.

Se encontraron 4 fases principales, en el flujo de datos que describen las técnicas de visualización encontradas:

1. Recuperación de datos de entrada: donde se establecen dos procesos principales: (1) la detección de opiniones, y (2) La extracción de información relevante (características, sentimientos y temas).
2. Detección de orientación o polaridad.
3. Resumen de la evaluación de orientación y polaridad.
4. Implementación de técnicas de visualización.

El estado del arte de las técnicas de análisis visual inicia en la Sección 2.1, con la definición de las estructuras de datos típicas en las que son almacenadas las opiniones en las plataformas de comercio electrónico. Luego, en la Sección 2.2 se exponen las principales técnicas para la extracción de información, cuyas principales vertientes se centran en la extracción de tópicos y características. A continuación, en la sección 2.3 se exponen las técnicas encontradas para la clasificación de opiniones, donde en la sección 2.4 se expande con mayor profundidad sobre las mismas. Este capítulo finaliza con la sección 2.5, que expone los tipos de técnicas de análisis visual encontrados.

2.1. Estructura de los datos de entrada

En lo que al análisis de las técnicas de visualización de opiniones incumbe, en la bibliografía estudiada se muestran diversidad de formatos y fuentes de datos, donde cada autor dispone de las mismas según el contexto del problema a resolver.

2.1.1. Fuente de datos

Las fuentes pueden tener orígenes distintos, algunos autores centran su estudio en fuentes particulares y otros en fuentes mixtas, donde las principales fuentes encontradas se listan a continuación:

1. Sitios web de comercio electrónico: Las plataformas web de comercio electrónico, ofrecen espacios a sus clientes donde pueden expresar su opinión. Allí elaboran una descripción en un formato de texto, donde adicionalmente se suelen responder unas series de preguntas asociadas a distintas dimensiones, con el fin de estructurar la información. Ejemplo del uso de este tipo de fuente se observa en [43] con su propuesta “Opinion Observer”.
2. Redes sociales: A pesar de que el comercio electrónico no es el eje principal de este tipo de plataformas, las redes sociales desarrollan un entorno de participación y libre opinión sobre productos y servicios. Los usuarios expresan sus opiniones sin seguir ningún formato preestablecido. Se puede encontrar información no estructurada en todas sus presentaciones: descripciones de texto libre, fotografías, audio y vídeo. En [39] analizan el fenómeno en redes sociales, donde centran su estudio en el análisis de opiniones enmarcado en el contexto de posturas predeterminadas.
3. Foros y blogs: Estas se posicionan como dos plataformas cuyo contenido es más especializado y las opiniones expresadas por los usuarios típicamente se escriben en torno a una temática en común. Este espacio a pesar de que permite compartir información de distinto tipo,

muestra una predominancia del formato de texto libre, ejemplo de esto es el trabajo desarrollado en [10].

4. Encuestas: muy utilizadas para obtener feedback de los clientes en marketing, como en el caso del sistema DemographicVis [18], en cuya implementación se utilizaron datos recuperados mediante encuestas digitales. Este tipo de formato suele favorecer el almacenamiento de datos estructurados, aun así existen formatos donde se permite incluir fragmentos de texto libre.

Los documentos recuperados de cada fuente se pueden clasificar en tres categorías, dependiendo del nivel de abstracción de los datos [63]:

1. A nivel de documento: El documento se considera como una entidad donde el análisis aplicado se hace a todo el formato.
2. A nivel de frase: La frase es considerada la entidad donde el análisis de cada unidad lingüística se hace al nivel de conjunto de palabras. Este proceso requiere una fase para resumir los resultados, de modo que se pueda obtener un resultado general de todo el documento, resumiendo un grupo de frases.
3. A nivel de aspecto: Realiza un análisis directo sobre las opiniones encontradas. Este nivel provee un análisis de sentimiento sobre los elementos de la opinión, por lo tanto requiere una fase previa para identificar la opinión y sus componentes.

2.1.2. Formatos de la fuente de datos

Según [43], los tipos de opinión según el tipo de plataforma muestran tres tipos de formato de entrada:

1. Formato 1: Las opiniones vienen estructuradas en positivas y negativas, de modo que la orientación de cada opinión no tiene que ser evaluada.
2. Formato 2: Las opiniones vienen en un formato libre, donde no se indica a priori la polaridad de la descripción dejada por el usuario.
3. Formato 3: Combina el formato 1 y 2, el usuario tiene la opción de responder en un formato libre los comentarios que desee aportar, pero también puede hacer una descripción aparte de sus opiniones con polaridad positiva y negativa respecto a un producto o servicio.

2.1.3. Complejidad de las opiniones

Las opiniones de usuarios tienen una alta diversidad mostrando diferentes formas de expresión. Los usuarios pueden hacer mención directa de adjetivos, sustantivos y verbos, para comunicar su opinión. En otras ocasiones puede hacer referencia a estos de forma indirecta, planteando dos categorías de opiniones [63]:

1. Aspectos Explícitos: Se hace mención explícita de los aspectos de opinión en la frase. Ejemplo: “El ordenador tiene una batería de larga duración”. Este ejemplo contiene una opinión explícita del atributo “batería”.
2. Aspectos implícitos: No están expresados explícitamente los aspectos de la opinión en la frase: Ejemplo: “Mi ordenador se descarga muy rápido”. Este ejemplo contiene una opinión implícita del componente “batería”, se refiere a él, pero nunca se menciona.

2.2. Extracción de información de las opiniones

Las técnicas de visualización deben de tener un eje central, sobre el cual presentar el gráfico. Dicho eje es determinado por la información que contiene, o se puede extraer de la fuente de datos. Así si se desea hacer énfasis en los sentimientos asociados a las opiniones, se debe implementar una rutina para extraer los sentimientos de las opiniones. El mismo caso sucede para las características y tópicos, se deben implementar algoritmos para la extracción de información que permitan contextualizar cada opinión.

2.2.1. Extracción de tópicos

Dado un grupo de documentos, se pretende asociar grupos de palabras que cohabitán juntas. Los grupos de palabras son identificados con técnicas de aprendizaje automático, los grupos de palabras encontrados se deben analizar, para etiquetar y comprender el contexto de cada grupo.

Existen múltiples técnicas para el análisis de tópicos, los basados en reglas como en [14] o modelos bayesianos como en [58]. En los trabajos de análisis de opiniones y visualización destaca con claridad LDA (Latent Dirichlet Allocation) sobre los métodos mencionados anteriormente e inclusive sobre otros métodos para modelar semánticas de palabras basadas en temas, como LSA¹ y PLSA² como en [16] y [48].

¹Latent Semantic Analysis: algoritmo que encuentra una representación de temas mediante documentos y palabras.

²Probabilistic Latent Semantic Analysis: Algoritmo que usa un enfoque probabilístico para encontrar palabras que definen un tópico.

LDA permite encontrar grupos de palabra similares en un documento, identificando categorías o tópicos latentes en corpus de documentos [5]. Bajo este enfoque se identifican bolsas de palabras que representan un grupo de tópicos en el documento. Allí no es tenido en cuenta el orden de las palabras, por lo que se considera el orden no aporta información relevante. Bastantes trabajos tienen como base LDA para la construcción de modelos de tópicos en el contexto de minería de opiniones, como [42], [51] y [102].

El flujo de proceso para la extracción de tópicos utilizando LDA se describe a continuación:

1. Lectura de conjuntos de datos.

2. Selección de espacios con opiniones para su análisis.

3. Pre procesamiento de texto:

 Tokenización.

 Eliminación de palabras vacías.

 Aplicación de algoritmo de Porter Stemmer³ [62].

 Construcción de diccionario de palabras.

 Filtrado de tokens de baja frecuencia.

 Construcción de Corpus.

4. Modelo LDA:

 Definición de parámetros del modelo.

 Entrenamiento del modelo.

 Aplicación de modelo sobre el conjunto de datos.

5. Volcado de información:

 Seleccionar tópicos de mayor probabilidad por cada opinión.

 Aregar una columna al conjunto de datos que incluya el tópico de mayor probabilidad en cada opinión.

 Guardar el conjunto de datos.

2.2.2. Extracción de características

La extracción de características adquiere relevancia para enfocar el proceso de análisis de sentimientos, permitiendo asociar un evaluación de una opinión con una característica de un producto o servicio. Las técnicas implementadas tendrán su aplicación dependiendo de la estructura misma de los datos, ya que algunos trabajos se enfocan en identificación de palabras asociadas a opiniones, para detectar posteriormente su orientación semántica, según [2] los métodos podrían clasificarse como se muestra a continuación:

³Método para reducir una palabra a su raíz o a un stem.

Métodos con base en aprendizaje automático: Destacan en cantidad trabajos orientados al aprendizaje automático donde aplican técnicas como reglas de asociación enfocándose en la identificación de frases con sustantivos de mayor frecuencia como en [28], [29] ó en [34] que incluye modelos ocultos de Markov. Bajo enfoque semi supervisado destaca el trabajo de [98] que realizan la extracción de características mediante el algoritmo EM (esperanza-maxificación) basado en probabilidad bayesiana. En esta misma categoría aparece un trabajo con un contexto más reciente [59] desarrollado en base a resúmenes estadísticos. Bajo enfoque no supervisado destaca LDA como en [7] que utiliza variaciones de esta técnicas para la extracción de propiedades en documentos de texto. En esta misma línea aparece de nuevo [98] que plantea comparaciones de su algoritmo EM con variaciones del algoritmo de agrupamiento K-medias.

Métodos con base en ontologías: Se emplea una estructura que establece relaciones de un grupo finito de palabras asociadas a un conjunto de productos, para identificar características en un grupo de opiniones de usuarios dadas. Algunos ejemplos aparecen en [22], [60] y [101].

Métodos con base en Lexicón: Dada una lista de palabras asociadas a los distintos elementos de productos, se construye un lexicón de características, para la identificación de las mismas en opiniones de usuarios [41].

Métodos con base en relación de dependencia: Se aplican reglas para la extracción de características, teniendo en cuenta las relaciones de dependencia de término, vistos en un conjunto de opiniones dado [64], [74], [93] y [100].

Numerosos trabajos describen el proceso de extracción de características en el análisis de opiniones, entre estos destaca [43], donde se plantea un enfoque basado en reglas descrito como sigue, donde N hace referencia a sustantivos, V a verbos, enumerándose estas por su orden de aparición:

1. POS taging de cada palabra de los documentos.
2. Remover dígitos numéricos.
3. Agrupar palabras que representen la mismas características (sinónimos).
4. Usar N-Grams para producir pequeñas secuencias de palabras.
5. Distinguir secuencias de palabras con etiquetados POS duplicadas en un mismo N-Gram.
6. Implementar algoritmo de Porter stemming.

7. Guardar secuencias de N-Grams como un conjunto de datos transaccional.

8. Implementación de reglas de asociación:

Minino soporte 1 porciento.

Reglas de la forma : N1,N2 entonces característica y V, entonces característica

9. Pos procesamiento:

Descartar todas las reglas que no contengan características en el consecuente.

Dividir reglas que puedan contener otras reglas en su interior (Como las reglas que contienen verbos, que pueden ser característica también).

Filtrar las reglas con confianza mínima del 50 porciento (eliminar reglas que no son lo suficientemente predictivas).

10. Generar patrones de lenguaje (N1,N2 entonces característica a N1 característica N2 por ejemplo).

11. Extracción de características (manejo de excepciones).

Aceptar huecos en los patrones buscados (ejemplo: buscando NN1 característica NN2 podría ser valido “size of printout”).

Añadir nuevo vocabulario de características encontrado en el proceso (por ende se añaden nuevos patrones de búsqueda).

En caso de empate de un segmento que satisface múltiples condiciones, se selecciona la de mayor confianza.

El sistema de reglas no es robusto a palabras solitarias, como “pesado” ó “grande”, por lo que este tipo de candidatas serán tratadas aparte.

Permitir en esta primera etapa la visualización de una secuencia larga de la frase encontrada para una mejor evaluación.

12. Corregir excepciones:

Conflictos con más de un candidato por frase.

Filtrar palabras fuera de contexto, con alta frecuencia de aparición que han sido valoradas como “características” por el sistema (ejemplo: se encuentra característica “hum” para la frase “slight hum from subwoofer when not in use”).

2.3. Clasificación de opiniones

Las opiniones a analizar necesitan ser evaluadas para determinar su orientación. Las técnicas de análisis de sentimientos, permiten realizar dicha evaluación con diferentes grados de granularidad. Algunas de las técnicas hacen énfasis en la clasificación de diferentes tipos de sentimientos, otros en analizar el grado de afinidad de los sentimientos (descubriendo su polaridad). A continuación se ilustran los distintos enfoques encontrados.

2.3.1. Ontologías

Con base en ontologías: Mediante estructuras de orden jerárquico se establecen relaciones entre los atributos de productos y sus propiedades, para la construcción de un clasificador de sentimientos, como en [80], [88] y [103].

2.3.2. Lexicón

Con base en lexicón: Estos enfoques realizan la evaluación de sentimientos asociados a comentarios, comparando palabras con un diccionario que permite identificar la opinión en cuestión. Los términos de las frases utilizados son palabras de sentimientos analizadas típicamente con granularidad a nivel documento o frase como en [3], [32], [43] y [78]. Estos trabajos calculan los sentimientos con operaciones aritméticas y estadísticas con diversos criterios, para determinar el valor de un sentimiento en una frase o documento. Un enfoque más complejo plantea el uso de técnicas de procesamiento del lenguaje natural (analizando sentimientos de adjetivos, verbos y sustantivos) para además extraer valor del componente semántico de las opiniones [37].

2.3.3. Aprendizaje automático

Con base en aprendizaje automático: El factor común aquí son las técnicas de aprendizaje supervisado, donde se encuentran trabajos que emplean para la clasificación de opiniones algoritmos como SVM (Máquinas de soporte vectorial) [35], clasificadores bayesianos⁴, máxima entropía [85] y redes neuronales [79]. Otros autores combinan varias de estas técnicas bajo enfoques híbridos que buscan explotar las bondades de cada una de estas técnicas en cada etapa, como los mencionados en [36].

2.3.4. Polaridad de opiniones

Dentro del contexto de visualización y minería de opiniones, se vuelve de gran interés identificar la polarización de las opiniones. Los enfoques

⁴Clasificador probabilístico fundamentado en el teorema de Bayes.

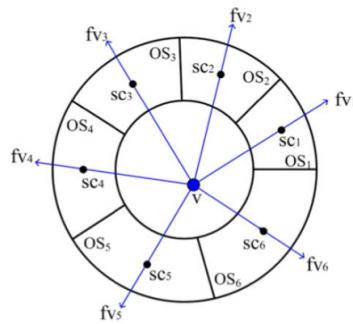


Figura 2.1: Opinion Ring: Distribución del espacio de opiniones [19].

encontrados plantean diferentes métricas que satisfacen a ciertas técnicas de visualización. Las técnicas de evaluación de polaridad, se enfocan en descubrir la tendencia de una opinión en una o varias dimensiones, donde se incluyen: positiva, negativa, neutral, con nivel de incertidumbre y de granularidad.

Evaluación de polaridad: positiva

Este método se estudia en [19], con su propuesta el Opinion Ring. Los autores se centran en el análisis de tendencia de las opiniones para favorecer un tipo de opinión u otra. Así conocida la problemática a analizar, se plantean los temas de opinión principales para evaluar cada una de las opiniones. Se trata de encontrar el nivel de pertenencia de esta a cada dimensión, que ha sido configurada de acuerdo a las temáticas de interés planteadas. Un enfoque adicional que plantea el Opinion Ring, consiste en identificar las opiniones que no tienen una tendencia definida, ubicando dichas opiniones en el centro del círculo. La opinión a evaluar estará ubicada cerca del tema con mayor probabilidad de pertenecer. Una opinión puede tener un grado de pertenencia parcial a otros temas, indicando una tendencia a poder pertenecer a otro. En la Figura 2.1 se muestra un espacio de opiniones: OS1,OS2,OS3,OS4,OS5 Y OS6, donde la opinión V, ubicada en el centro del gráfico, muestra un estado actual pasivo hacia alguna de las opiniones, este tiene una tendencia positiva hacia la opinión OS6 y otra ligeramente menor hacia OS3.

Evaluación de polaridad: positiva y negativa

Bajo este enfoque se plantea una evaluación de la polaridad de cada opinión en dos dimensiones, para decidir si una opinión tiene inclinación positiva o negativa, bajo este primer planteamiento la medida no tiene ningún

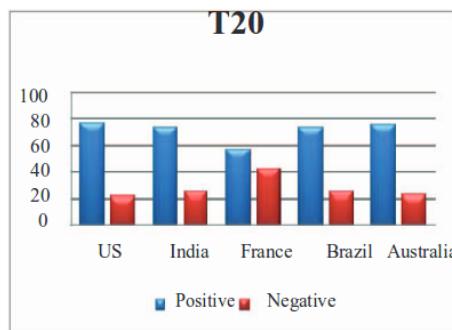


Figura 2.2: Diagrama de barras: porcentaje de sentimientos por país sobre T20 [54].

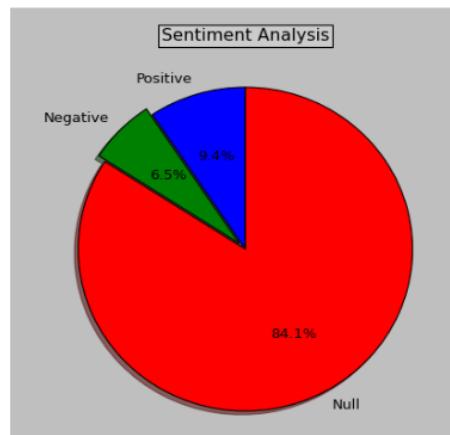


Figura 2.3: Diagrama de Pie: Proporción de sentimientos [24].

nivel de sensibilidad y se desconoce el grado de pertenencia hacia cualquiera de las orientaciones. Los métodos de visualización implementados bajo este enfoque hacen énfasis en la cantidad de opiniones positivas y negativas como se puede observar en [54] y en [24], donde se representa en ambas la magnitud de la orientación de las opiniones, como se muestra en la Figura 2.2 y en la Figura 2.3.

Evaluación de polaridad: positiva y negativa con granularidad

Ciertos problemas de negocio se pueden resolver conociendo únicamente los cambios de polaridad u orientación de las opiniones. En algunos casos se requiere identificar el grado o nivel de la orientación, para poder realizar comparaciones entre opiniones de acuerdo al nivel de la orientación de las mismas. Dentro de este tipo de categorización encajan un gran número de

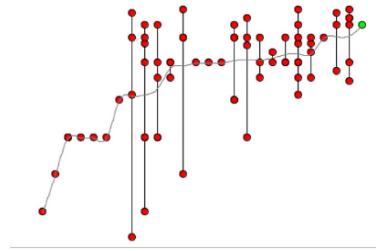


Figura 2.4: Ejemplo del mapeo gráfico: SentiVis[17].

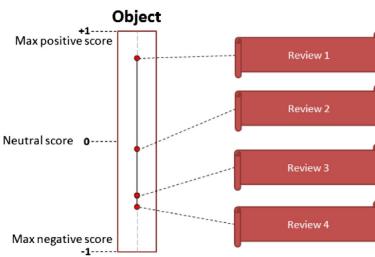


Figura 2.5: Mapeo gráfico: SentiVis[17].

visualizaciones. Ejemplo de esto se puede observar en [17] con su propuesta denominada SentiVis, donde como se observa en la Figura 2.4, se representa con un mapeo gráfico de líneas verticales cada objeto de análisis, los puntos en cada línea vertical hacen referencia a las puntuaciones realizadas sobre cada opinión. Los puntos de mayor altura representan puntuaciones más altas o positivas, y los de menor altura se refieren a puntuaciones más bajas o negativas de cada opinión, como en el ejemplo de la Figura 2.5.

Evaluación de polaridad: positiva, negativa y neutral

Muchas de las opiniones realmente no tienen un nivel significativo de tendencia hacia las orientaciones positivas y negativas. Este tipo de opiniones puede contener contenido neutro, lo que propone este enfoque es identificar y evaluar las opiniones, teniendo en cuenta esta nueva dimensión. El objetivo es evitar que opiniones de naturaleza neutral sean evaluadas como positivas o negativas. En [84] se ilustra un método que comparte este enfoque, en la Figura 2.6 se visualiza un diagrama denominado mapa de relaciones. El área de cada elipse corresponde al total de comentarios sobre un tema, los puntos representan los distintos participantes o usuarios, el tamaño de los puntos hace referencia al número de palabras con sentimientos asociados a la orientación de la opinión, (donde un usuario puede tener mas de una opinión sobre el tema) y la línea en verde conecta usuarios con intereses comunes.

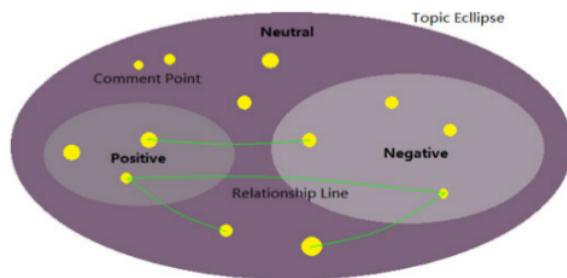


Figura 2.6: Mapa de relaciones [84].

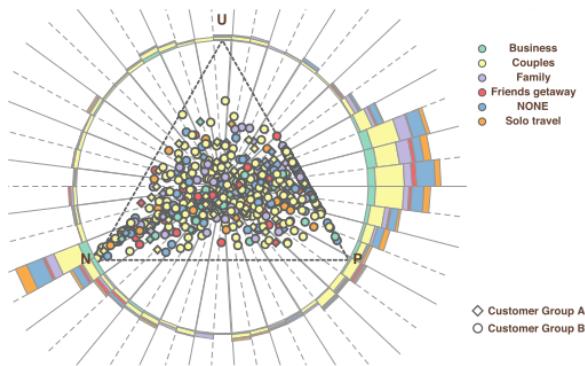


Figura 2.7: Opinion Seer: Comparación visual de dos grupos de clientes A y B [91].

Evaluación de polaridad: positiva, negativa e incertidumbre

Los métodos empleados para evaluar la orientación de una opinión pueden ser poco sensibles para identificar o ponderar la certeza de dicha evaluación, el Opinion Seer [91] plantea un enfoque que incluye una nueva dimensión, la incertidumbre. Los autores desarrollan una aplicación de este concepto donde se emplea la simetría de un círculo para representar el espacio de opiniones. Encima de este se sobreponen una geometría triangular para representar las tres dimensiones que ponderan la orientación de los sentimientos (positivo, negativo e incertidumbre), como se le ilustra en la Figura 2.7.

2.4. Técnicas utilizadas en minería de opiniones

Una vez se ha identificado la polaridad de las opiniones, se hace necesario agrupar y resumir dicha información. Los métodos de evaluación de polaridad muestran un rango y tipo de variable distinto según el criterio de

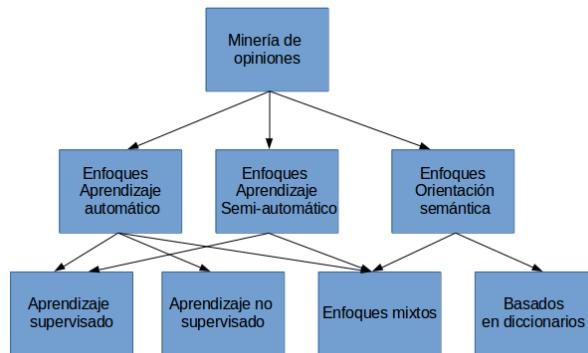


Figura 2.8: Enfoques de técnicas: minería de opiniones [63].

evaluación empleado. Se debe de incluir para cada uno de estos, una técnica que resuma la información de modo que esta pueda ser visualizada.

El análisis puede ir enfocado con varios niveles de granularidad, donde según [77] los que destacan son:

1. Minería de opiniones a nivel de documento.
2. Minería de opiniones a nivel de oración.
3. Minería de opiniones granularidad fina.

Las técnicas y herramientas que se muestran a continuación se sitúan dentro del contexto de aprendizaje automático y semi automático; supervisado y no supervisado. También se emplean técnicas basadas en métodos de orientación semántica [63]. Así, si se dispone de datos con etiquetas o de los recursos para realizar el etiquetado, una técnica de clasificación de aprendizaje automático resolvería el problema de evaluación de sentimientos. Si no se dispone de los recursos antes mencionados, se suele resolver el problema bajo un enfoque no supervisado o bajo el apoyo de técnicas de orientación semántica como las herramientas de tipo lexicón de sentimientos. A la larga estas técnicas se pueden combinar como en [43] donde se implementa un enfoque mixto con una implementación propia de un análisis sintáctico que evalúa los sentimientos con un lexicón de sentimientos, aplicando sobre estos técnicas de reglas de asociación. Los enfoques descritos se muestran en la Figura 2.8, complementando la descripción de [63] con los nuevos enfoques encontrados.

Dentro del tipo de técnicas mencionadas destacan los métodos semi automáticos (para el caso de aprendizaje automático supervisado y no supervisado) como en [43], donde se diseña una interfaz de usuario para evaluar instancias difíciles de clasificar, logrando excelentes resultados.

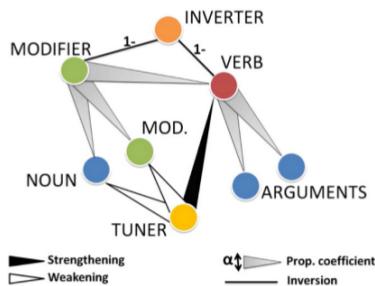


Figura 2.9: Arquitectura general de propagación entre componentes lingüísticos [17].

2.4.1. NPL: Para extracción de conocimientos y evaluación de la polaridad

El procesamiento de lenguaje natural es ampliamente utilizado en lo que procesamiento y evaluación de sentimientos concierne. En el ámbito de visualización de opiniones destaca una técnica muy utilizada, el análisis sintáctico mediante dependencias parciales. Los elementos lingüísticos se relacionan como se muestra en la Figura 2.9, donde cada elemento se interpreta así:

1. MODS: Son todos los constructores que modifican un verbo, sustantivo o otro modificador. Ellos pueden ser simples adjetivos, sustantivos o adverbios.
2. TUNS: Es un tipo de adverbio modificador que fortalece (o debilita) el valor de la palabra que expresa el sentimiento. Por ejemplo el adverbio “muy” aumenta tanto positiva como negativamente el valor de una palabra, dependiendo del sentimiento que exprese.
3. INVS: Los inversores vienen de la negación y son representados por palabras como “no” y “nunca”.
4. PREPS: Las preposiciones “a”, “con”, “en” y “así” son consideradas como canales que conducen el sentimiento entre términos que ellos conectan.
5. VERBS: Los verbos pueden tener connotaciones de sentimientos de todos los valores que transmiten a sus argumentos.

Una aplicación de NPL bajo el enfoque de análisis sintáctico, es la implementación de la universidad de Stanford NLP parser [76] para análisis de sentimientos, como se muestra en [17] y [72]. El análisis sintáctico aprovecha la estructura y el orden de verbos, modificadores, sustantivos y adjetivos,

Basic Dependencies	Dependencies for Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas
	<pre> nsubjpass(submitted, Bills) auxpass(submitted, were) agent(submitted, Brownback) nn(Brownback, Senator) appos(Brownback, Republican) prep_of(Republican, Kansas) prep_on(Bills, ports) conj_and(ports, immigration) prep_on(Bills, immigration) </pre>

Figura 2.10: Representaciones gramaticales entre palabras en una frase [72].

para construir una representación de dependencias con las relaciones gramaticales mediante un enfoque probabilístico, como se ilustra en el ejemplo de la Figura 2.10.

En el contexto de análisis de opiniones aparece otra interesante aplicación del procesamiento de lenguaje natural para identificar y extraer características en el espacio de opiniones. Entre estos se destacan dos trabajos: [43] y [28], que aprovechan la estructura sintáctica de los n-gram de cada opinión para identificar características explícitas en los textos. La estrategia de extracción mostró niveles de precisión superiores al 80 por ciento, pero aun plantea grandes retos para detectar características implícitas en los textos. La estrategia propuesta emplea reglas se asociación que obtiene reglas de la siguiente forma:

$$< N1 >, < N2 > - > [feature] \quad (2.1)$$

$$< V >, easy, to- > [feature] \quad (2.2)$$

$$< N1 > - > [feature], < N2 > \quad (2.3)$$

$$< N1 >, [feature] - > < N2 > \quad (2.4)$$

El autor emplea un procesador NPL [55] que reconoce las estructuras gramaticales para identificar verbos “V” y sustantivos “N” para las extracción de características según la posición que describan los mismos. Posteriormente se identifica que las únicas reglas útiles son las de la forma (1) y (2), por lo que (3) y (4) son descartadas. El entrenamiento del modelo se realiza bajo un enfoque supervisado donde varias de las características en las opiniones han sido a priori etiquetadas manualmente, buscando que el sistema se pueda generalizar a nuevas opiniones.

Otro ejemplo relevante en la literatura en el análisis de opiniones para la extracción de características es el modelado de Popescu and Etzioni [61] que estudia las interacciones de los temas de interés a través de un árbol de dependencia de características. Estos mismos autores plantean una estructura novedosa donde tienen en cuenta para la evaluación de la polaridad de una

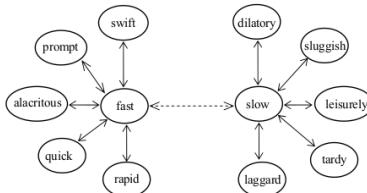


Figura 2.11: Estructura de adjetivos bipolares [28].

palabra un enfoque colectivo teniendo en cuenta la estructura de conectores en la frase. Otros trabajos destacables en contexto de minería de reviews son [28], [38], [52],[53],[95] y [96].

2.4.2. Sentiment lexicon: para evaluación de la polaridad de las frases

Existe una variedad de plataformas digitales en múltiples idiomas que ponen a su disposición diccionarios que permiten evaluar la orientación de palabras, entre estas se encuentran: ANEW [6], WordNet [50] y wiktionary [97] por nombrar algunas. Esta herramienta es explotada con distintas estrategias, por ejemplo [28] la emplea para poder identificar la polaridad de los adjetivos, haciendo uso de los posibles antónimos y sinónimos del mismo identificando así una orientación de la opinión. Ejemplo de esto se ilustra en la Figura 2.11, donde el conjunto de palabras de la izquierda hace referencia al adjetivo “rápido” y las posibles formas del mismo y a la derecha el adjetivo “lento” con sus posibles sinónimos.

Algunos trabajos plantean la elaboración de sus propios diccionarios como es el caso de [30], que además incluye en su desarrollo un enfoque mixto que incorpora PLN y sistemas difusos en su propuesta. Otro trabajo relevante [43], plantea una solución mixta empleando diccionarios digitales de las plataformas antes mencionadas y un diccionario propio. Como primera instancia los adjetivos a evaluar se buscan en el diccionario local, de no encontrarse el sistema realiza una búsqueda y agrega esta nueva definición al diccionario local. El pseudo código se describe en la Figura 2.12.

2.4.3. Algoritmos de clasificación: Para extracción de frases de interés y evaluación de sentimientos

Utilizados extensamente en el contexto de análisis de opiniones, se utilizan principalmente bajo dos enfoques: el primero para identificar elementos de interés en el espacio de documentos de entrada como en [43], donde se usan reglas de asociación y umbrales de soporte y confianza para identificar opiniones y sus características. El segundo emplea técnicas de aprendizaje automático para clasificación, con el objetivo de identificar la orientación de

```

1. Procedure OrientationSearch(adjective_list, seed_list)
2. begin
3.   for each adjective wi in adjective_list
4.   begin
5.     if (wi has synonym s in seed_list)
6.       { wi's orientation= s's orientation;
7.       add wi with orientation to seed_list; }
8.     else if (wi has antonym a in seed_list)
9.       { wi's orientation = opposite orientation of a's
          orientation;
10.      add wi with orientation to seed_list; }
11.   endfor;
12. end

```

Figura 2.12: Pseudo código buscador de orientación de palabras [28].

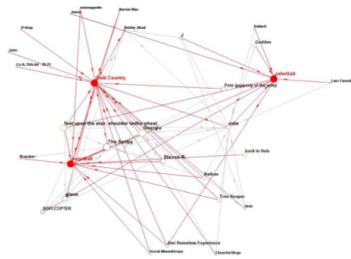


Figura 2.13: Red social del agrupamiento sobre el tema Venezuela apaga RCTV [94].

las opiniones, un ejemplo de este enfoque es [17], que emplea simultáneamente técnicas de PLN.

En la búsqueda bibliográfica realizada destaca la implementación de algoritmos como: Apriori en [43], Naive Bayes y maquina de soporte vectorial en [87] que realiza una comparación de ambas técnicas.

2.4.4. Algoritmos de agrupamiento: para el análisis de interacciones

Muchos analistas centran su interés en entender las interacciones y relaciones entre las opiniones de los usuarios, para identificar usuarios con opiniones similares respecto a un tema, como lo plantean en [94]. En este trabajo se comparan técnicas de agrupamiento basadas en densidad, como DBSCAN y SDC. Como aplicación en técnicas de visualización, estos mismos autores utilizan una red de grafos con aristas con una dirección, para indicar el tipo de interacción entre los usuarios. Este método de visualización se ilustra en la Figura 2.13, el gráfico describe una red destacando sus elementos mas representativos, aun así con redes de usuarios muy grandes la gráfica se puede saturar fácilmente como se verá mas adelante.

2.4.5. LDA para la extracción de tópicos

La extracción de tópicos adquiere interés en el la minería de opiniones, bajo la perspectiva de proporcionar un grado de granularidad que muestre los distintos temas de los que se hace mención en cada comentario. Para esto [5] propone un modelo para documentos de texto que genera un grupo de tópicos “k”, que es representado por una distribución multinomial sobre “V” palabras en el vocabulario. El algoritmo toma una mezcla de muestras para generar sus resultados.

Los resultados consisten en un documento de “N” palabras $w = \{w_1, \dots, w_N\}$ que es generado a través de procesos iterativos que tienen como base la distribución Dirichlet⁵, como se muestra continuación.

Una representación del algoritmo se puede observar en la Figura 2.14 donde el proceso sigue los pasos mostrados:

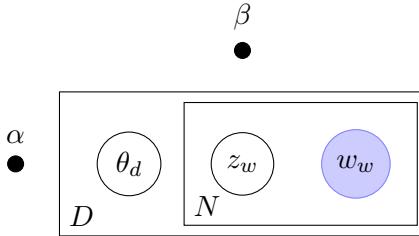


Figura 2.14: Diagrama de LDA [45].

- 1: **for** documento d_d in corpus D **do**
- 2: seleccionar $\theta_d \sim \text{Dirichlet}(\alpha)$
- 3: **for** posición w in d_d **do**
- 4: Seleccionar un tema $z_w \sim \text{Multinomial}(\theta_d)$
- 5: Seleccionar una palabra w_w from $p(w_w|z_w, \beta)$, una distribución multinomial sobre las palabras condicionadas al tema y al anterior β .
- 6: **end for**
- 7: **end for**

En temas de minería de opiniones LDA es ampliamente utilizado, donde destacan trabajos como [51] y [99]. En cuanto temas de visualización empleando LDA destaca el proyecto LDAvis [71] cuya librería se puede descargar en el siguiente enlace <https://github.com/cpsievert/LDAvis>.

2.5. Técnicas de visualización

La idea general de cualquier tipo de técnica de visualización es transmitir la información en un gráfico, donde según su topología puede ser evaluada según las métricas y áreas de la Tabla 2.1. El planteamiento es desarrollado

⁵Familia de distribuciones de probabilidad continuas multivariadas

Assesment area	Metrics
Visual impact	Eye pelasing
Overall performance	Easy to understand, User-friendly
Overall design style	Informative, Intuitive
Information quality	Userfulness, Comprehensiveness
Visual representation model	Comparasion ability, representation style
Information presentation model	Pre-knowledge required

Tabla 2.1: Áreas y métricas para la evaluación de técnicas de visualización.

Visualizations	Metric	Data collection method	Mean	SD	t value	p value
Bar chart	Comparison	Seminar	3.42	1.21	-2.357	.021
		Online questionnaire	3.89	.979		
Line graph and pie chart	Easy to understand	Seminar	3.40	1.32	-4.184	.000
		Online questionnaire	4.19	.856		
	User-friendly	Seminar	3.30	1.19	-4.372	.000
		Online questionnaire	4.14	.931		
	Informative design	Seminar	3.26	1.19	-2.874	.005
		Online questionnaire	3.78	.832		
	Usefulness	Seminar	3.17	1.20	-4.666	.000
		Online questionnaire	3.94	.715		
	Comprehensiveness	Seminar	3.03	1.28	-3.307	.001
		Online questionnaire	3.64	.833		
	Comparison	Seminar	2.84	1.19	-4.598	.000
		Online questionnaire	3.64	.798		
	Representation style	Seminar	2.92	1.13	-4.321	.000
		Online questionnaire	3.56	.607		
Bar chart with symbols	Comprehensiveness	Seminar	2.73	1.17	-2.259	.027
Glowing bars	Eye pleasing	Online questionnaire	3.14	.867		
		Seminar	4.10	1.04	4.477	.000
	Easy to understand	Online questionnaire	3.22	.959		
		Seminar	3.85	1.28	2.028	.044
	Informative design	Online questionnaire	3.36	1.13		
		Seminar	3.80	1.16	2.023	.045
		Online questionnaire	3.36	1.02		

Figura 2.15: Evaluación de métodos de visualización parte 1 [70].

por [70], contiene 6 ejes fundamentales y son expuestos así: el área de impacto visual: responde a la pregunta ¿Es agradable a la vista? Rendimiento general: ¿Es fácil de entender y amigable con el usuario? Estilo de diseño general: ¿Es informativa e intuitiva? Calidad de la información: ¿Es útil y comprensiva? Modelo de representación visual: evalúa el nivel de habilidad de comparación y estilo de representación. Por ultimó esta el modelo de presentación de la información: cuestiona si es requerido cierto conocimiento previo, para poder comprender la visualización.

Los autores de [70] realizan una ponderación de las métricas para cada tipo de gráfico. Un grupo de expertos y no expertos realizo una evaluación de los tipos de visualización, como se muestra en las tablas de las Figuras, 2.15, 2.16 y 2.17 en cuyo caso se compara métodos gráficos: Bar char, Line graph and pie chart, Bar chart with symbols, Glowing bars, tree map, coordinated graph, positioning map, comparative relation map y visual summary.

Las evaluaciones realizadas en [70] muestran tipos de gráficos como Line graph y pie char muy intuitivos y fáciles de entender. Bar Char y Positioning Map se destaca por su capacidad para comparar datos. Otros se muestran más complejos y exhiben un nivel alto en la métrica “Pre-Knowledge requi-

Visualizations	Metric	Data collection method	Mean	SD	t value	p value
Coordinated graph	Intuitive design	Seminar	1.98	1.09	-2.448	.016
	Usefulness	Online questionnaire	2.50	1.13		
	Seminar	2.12	1.16	-2.848	.005	
	Online questionnaire	2.75	1.13			
	Comprehensiveness	Seminar	2.05	1.17	-3.215	.002
	Online questionnaire	2.75	1.05			
	Representation style	Seminar	2.15	1.30	-2.142	.035
	Online questionnaire	2.56	.877			
	Pre-knowledge required	Seminar	3.22	1.62	-3.408	.001
	Online questionnaire	4.03	1.08			
Positioning map	Eye pleasing	Seminar	2.43	1.19	-2.695	.008
	Online questionnaire	3.03	1.06			
	Seminar	2.58	1.30	-4.272	.000	
	Online questionnaire	3.50	1.06			
	User-friendly	Seminar	2.66	1.14	-3.554	.001
	Online questionnaire	3.42	.996			
	Informative design	Seminar	2.88	1.23	-2.715	.008
	Online questionnaire	3.39	.87			
	Intuitive design	Seminar	2.26	1.12	-2.575	.011
	Online questionnaire	2.81	1.01			
Comparative relation map	Usefulness	Seminar	2.65	1.28	-4.106	.000
	Online questionnaire	3.44	.909			
	Comprehensiveness	Seminar	2.53	1.11	-2.824	.005
	Online questionnaire	3.11	.979			
	Comparison	Seminar	2.49	1.16	-3.076	.003
	Online questionnaire	3.17	1.08			
	Representation style	Seminar	2.66	1.19	-2.260	.025
	Online questionnaire	3.17	1.06			
	Pre-knowledge required	Seminar	2.63	1.42	-2.932	.005
	Online questionnaire	3.33	1.20			
Representation style	Pre-knowledge required	Seminar	2.84	1.30	-2.119	.037
	Online questionnaire	3.28	1.00			
Pre-knowledge required	Seminar	2.79	1.27	-3.854	.000	
	Online questionnaire	3.61	1.05			

Figura 2.16: Evaluación de métodos de visualización parte 2 [70].

Visualizations	Metric	Data collection method	Mean	SD	t value	p value
Tree map	User-friendly	Seminar	3.63	1.20	1.955	.052
	Pre-knowledge required	Online questionnaire	3.19	1.01		
	Seminar	2.69	1.28	-3.931	.000	
	Online questionnaire	3.61	1.02			
Visual summary	Eye pleasing	Seminar	2.99	1.21	-2.149	.033
	Online questionnaire	3.47	1.03			
	User-friendly	Seminar	2.75	1.22	-2.133	.035
	Online questionnaire	3.22	.989			
	Intuitive design	Seminar	2.57	1.22	-3.890	.000
	Online questionnaire	3.22	.722			
	Usefulness	Seminar	2.89	1.16	-3.398	.001
	Online questionnaire	3.42	.649			
	Comprehensiveness	Seminar	2.80	1.03	-5.778	.000
	Online questionnaire	3.64	.639			
Comparison	Comparison	Seminar	2.76	1.17	-4.434	.000
	Online questionnaire	3.50	.737			
	Representation style	Seminar	2.83	1.28	-3.170	.002
	Online questionnaire	3.44	.909			
Pre-knowledge required	Pre-knowledge required	Seminar	2.79	1.34	-4.446	.000
	Online questionnaire	3.69	.95			

Figura 2.17: Evaluación de métodos de visualización parte 3 [70].

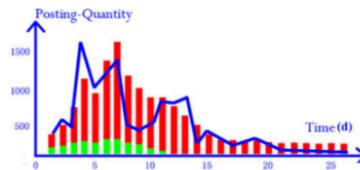


Figura 2.18: Número de publicaciones simuladas (línea) y datos reales (Barras) [84].

red” como es el caso de “coordinaed graph”, posicion map y visual summary. De esta manera se puede observar que cada tipo de gráfico exhibe unas propiedades distintas con evidentes ventajas para resolver problemas de negocio. Aquí se pueden usar tipos de gráfica con componente “easy to understand” sin la componente “pre-knoledge required”, para ilustrar usuarios no expertos, como los “Glowing bars”. Se puede plantear un enfoque totalmente distinto aprovechando la experiencia y conocimiento del usuario en el tema, para sacar provecho de la complejidad y capacidad de representación de la información del gráfico, como es el caso del “comparative relation map” o “Coordinated graph”.

Pero las métricas del gráfico no son el único criterio de decisión a tener en cuenta en el momento de preferir un tipo de visualización sobre otra. Hay propiedades intrínsecas del problema que como veremos a continuación, nos llevarán a preferir ciertos tipos de visualización según el contexto de los datos y el enfoque que se quiera dar.

2.5.1. Análisis visual temporal

Cuando se hace referencia al tiempo, este abarca todos los distintos formatos de meta datos que almacenen información de hora, día, mes año y sus posibles múltiplos y submúltiplos. La variable tiempo convierte secuencias de datos en series temporales que pueden ser analizadas como tal, no solo para comprender el pasado sino para predecir el futuro como en [84]. En la Figura 2.18 se visualiza el resultado de este trabajo, allí se presenta la evolución de los sentimientos positivos y negativos entorno a la de política de cero uso de papel en china. Las barras representan los datos reales, donde las de color verdes señalan las opiniones en contra y las rojas a favor de la política, la línea azul muestra la cantidad total de opiniones simulada.

Otro tipo de visualización relevante encontrada es la propuesta de [11], con un sistema de vistas coordinadas, donde se busca comprender la interacción de opiniones en el tiempo. En la Figura 2.19 se muestran las interacciones, donde los arcos sobre el eje horizontal representan las opiniones positivas y los que se encuentran debajo del mismo eje hacen referencia a opiniones negativas. El espesor de los arcos descritos representan el núme-

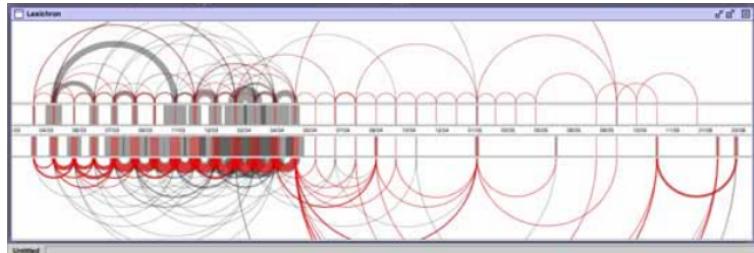


Figura 2.19: Vistas coordinadas representando interacciones de términos positivos y negativos [11].



Figura 2.20: Términos extraídos de opiniones positivas y negativas mensualmente [11].

ro de usuarios. Este tipo de visualización permite identificar patrones en el tiempo donde cada arco conecta dos puntos en diferentes momentos de tiempo que tienen términos en común en sus opiniones.

En [11] se plantea una visualización de términos positivos y negativos como una serie de tiempo, tal cual se muestra en la Figura 2.20. Los términos positivos y negativos son representados en gráficos separados donde cada columna representa los términos que aparecen cada mes.

Otro trabajo relevante es la propuesta de [90], el Opinion Flow, este ilustra la difusión y transición entre diferentes temáticas combinando la visualización planteada por Sankey [67], con la técnica de estimación de densidad de kernel (KDE). El desarrollo planteado muestra el flujo de opiniones entre usuarios y el flujo de usuarios a través de temas. La interfaz de usuario de la propuesta permite: seleccionar usuarios para examinarlos en detalle, interacción con el árbol de visualización de temáticas, ver el rastro de la influencia de usuarios sobre la difusión de opiniones, navegar en múltiples escalas de tiempo, examinar el comportamiento de la difusión de usuarios y eliminar temáticas y transiciones poco relevantes. La Figura 2.21 muestra una visua-

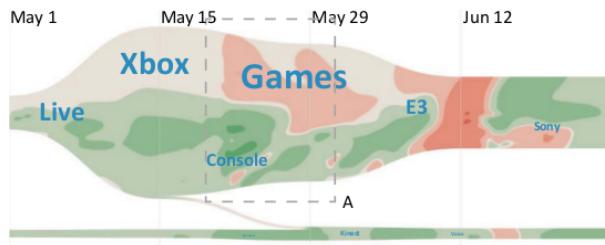


Figura 2.21: Difusión de opiniones de una consola de video juegos en el periodo comprendido entre mayo 15 al 29, cuando una versión del dispositivo fue anunciada [15].

lización de la interfaz de usuario con información extraída de redes sociales en el periodo de lanzamiento de una consola de video juegos. La zona A de la misma Figura muestra una zona en rojo para la palabra “Games”, que representa opiniones negativas causadas por cierta incompatibilidad del dispositivo. Las zonas del mapa verde y azul representan sentimientos hacia los temas positivos y negativos respectivamente. El eje horizontal comprende la dimensión clave en este análisis visual, puesto que muestra en el tiempo la evolución de las temáticas tratadas en la red social.

El trabajo anterior ha influenciado en gran medida diversas propuestas que usan como base del modelo visual el flujo de transición mostrado, un ejemplo de estos es [15], que incorpora nubes de palabras de los tópicos relevantes en zonas puntuales del gráfico.

En su desarrollo [12] plantea un análisis visual enfocado en tres aspectos principales: el primero identificar la estructura general conociendo la secuencia de temáticas de discusión, la agregación de opiniones en múltiples niveles estructurales de discusión y lo mas importante un análisis visual orientado en el tiempo. La Figura 2.22 muestra el gráfico de la propuesta donde al interior de un círculo se representa los temas de interés, la herramienta dispone de una interfaz que permite al usuario hacer clic sobre alguno de los temas de interés para conocer su comportamiento en el tiempo. Los círculos al interior del círculo principal están conectado con las temáticas de análisis, así cada una de estas áreas circulares representa los sub temas del tema principal con el que están conectados. La herramienta permite a los usuarios realizar cinco tipos de acciones: explorar temas y contenidos, explorar usuarios, explorar publicaciones relevantes de opiniones (de usuarios que empezaron un discusión o lideraron una temática), explorar la dimensión temporal y realizar filtrado y ordenación para deducir la complejidad del gráfico. La intensidad de los colores dentro del círculo representa la intensidad con que los usuarios apoyan el tema de referencia, donde colores más oscuros representan mayor apoyo en proporción de las publicaciones positivas o negativas realizadas. La Figura 2.22 contiene datos de discusión de un usuario sobre el tema del

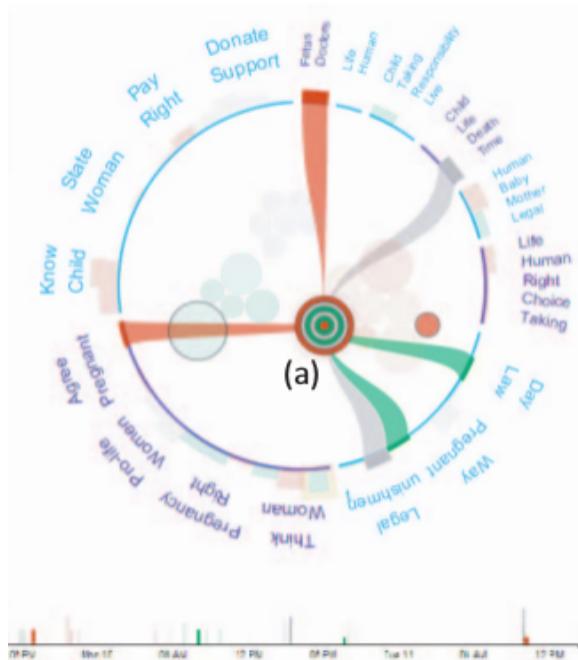


Figura 2.22: Exploración de la burbuja de un usuario en el hilo de opiniones [12].

aborto, allí se muestra que una de sus primeras opiniones negativas fueron bajo los temas “Fetus” y “Doctors” luego parece más adelante que el pensamiento del usuario cambia radicalmente, mostrando opiniones positivas hacia los temas “law”, “legal” y “pregnant”.

2.5.2. Análisis visual orientado a la comparación de temáticas:

La detección de temas es otro de los aspectos ampliamente estudiados. Se distinguieron dos enfoques principales, que dan solución a problemas de negocio distintos. El primer enfoque requiere un conocimiento previo del problema, donde expertos definen los temas de interés sobre los que se hace la búsqueda. Un ejemplo de esto es el Opinion Ring, el conjunto de entrenamiento empleado para el modelo recibe un etiquetado de los temas de interés, el resto se etiquetan como neutros mostrando un gráfico como el de la Figura 2.23, que analiza seis temas de interés: Genetic Algorithm, Probabilistic Methods, Theory, CaseBase, Reinforcement Learning y Rule Learning. Se busca con este método identificar las opiniones relacionadas con los temas de interés. En el caso de que las opiniones no pertenezcan a ningún tema, se buscará predecir la tendencia hacia alguno de los temas,

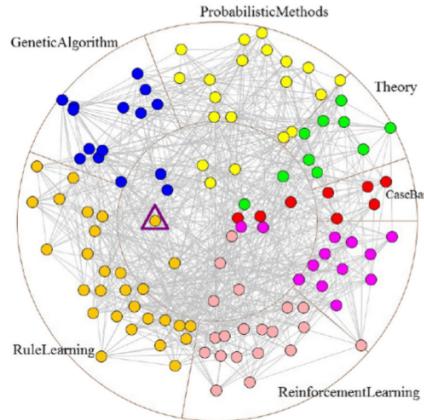


Figura 2.23: Visualización del Opinion Ring [19].

como es el caso de la opinión señalada con un triángulo en la Figura 2.23, donde la opinión tiene una tendencia hacia el tema Rule Learning.

El segundo enfoque tiene una etapa previa para detección del atributo de interés, donde puede ser sentimientos, características o los temas de las opiniones como en [23] con su prototipo “Pulse”. El sistema utiliza métodos de agrupamiento para encontrar los temas centrales de cada clúster, una vez son identificados se evalúa la polaridad de sus sentimientos y se representa mediante un tree map como en la Figura 2.24.

2.5.3. Análisis visual orientado a la comparación de productos y sus características

La detección de características o atributos mencionada en la sección 3 permite generar interesantes visualizaciones para comparar productos y sus características como en [43] con su propuesta “Opinion Observer”. Este trabajo realiza una comparación de los atributos de productos, mediante la extracción de sus características y evaluación de las mismas según las opiniones de los usuarios. El método de visualización se muestra en la Figura 2.25, donde se compraran cinco características de dos productos. Cada barra en el gráfico representa una característica, su tamaño la cantidad de opiniones que lo componen, así su ubicación determina la orientación de la opinión. El eje horizontal sirve como punto de referencia para dividir el espacio de opiniones positivas (hacia arriba) y negativas (hacia abajo).

Otro trabajo relevante se muestra en [92], donde son los indicadores de comparación como: “in contrast to”, “unlike”, “compare with”, “compare to”, “beat”, “win”, “exceed”, “outperform”, “prefer”, “than”, “as”, “same”, “similar”, “superior to”, “improvement over”, “better”, “worse”, “best”, “worst”, “more”, “most”, “less”, “least” son tenidos en cuenta para apren-

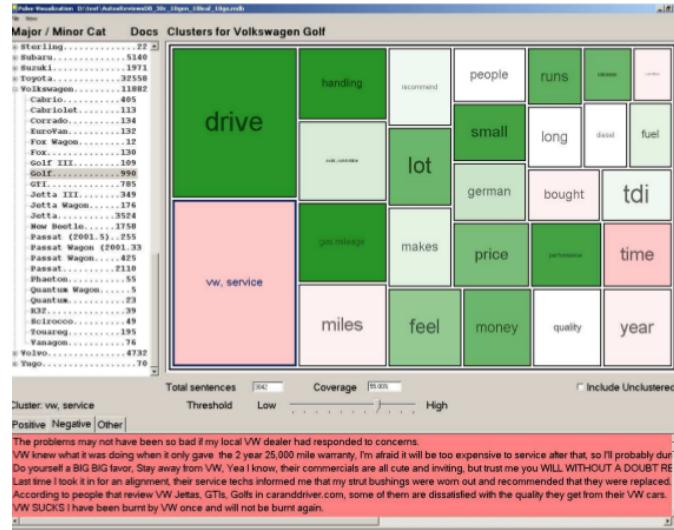


Figura 2.24: Captura de pantalla de la interfaz de usuario del prototipo “User”. Las palabras representan a los agrupamientos encontrados y el color los sentimientos [23].

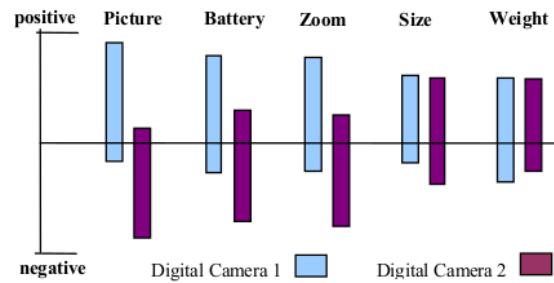


Figura 2.25: Visualización del Opinion Observer [43].

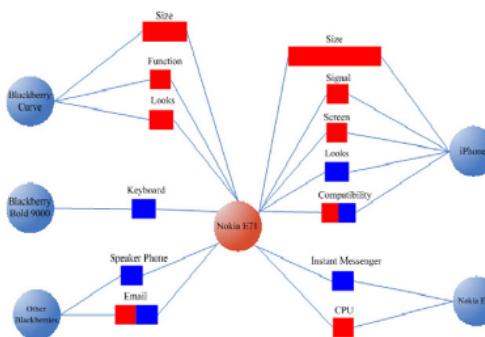


Figura 2.26: Mapa de relación competitiva [92].

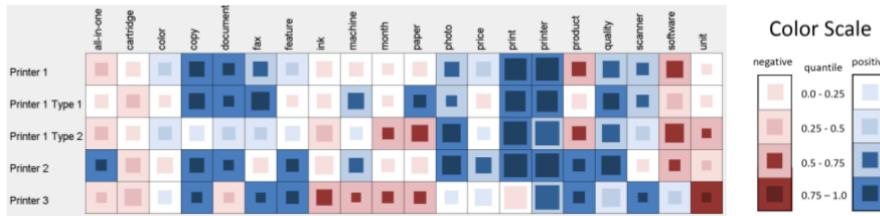


Figura 2.27: Tree Map: Resumen de reporte de impresoras [56].

der reglas mediante un clasificador que emplea SVM y redes bayesianas para identificar los patrones de comparación. El resultado se muestra en la Figura 2.26, donde las características comunes a los productos son identificadas y visualizadas con un grafo. Los círculos corresponden a los productos y los rectángulos a las características que los unen, el color dentro del rectángulo representa la característica favorecida por el producto de ese mismo color.

En la evaluación de los métodos de visualización mostrada en la Sección 2.5, destacó el Tree Map. Siguiendo esta misma línea, un trabajo relevante en el tema de análisis de reviews muestra una interesante aplicación para extraer atributos. La técnica consigue extraer las características más representativas, mediante TFIDF (Term Frequency Inverse Class Frequency) en pro de comparar productos y sus características [56]. El modelo plantea una representación en dos dimensiones donde en un eje se posicionan los productos y en el otro sus características, como en la Figura 2.27. El color en cada tupla producto-característica es representado por dos colores, el rojo para opiniones negativas y el azul para positivos. El tono del color indica el cuartil al que pertenece dicha tupla. El tamaño del cuadro interior representa la cantidad de clientes que han comentado dicho atributo.

En [84] se propone una interesante visualización con la implementación

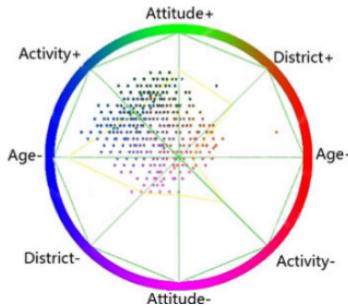


Figura 2.28: Sentiview: Atribute astrolabe [84].

de un astrolabio, donde en este se hace énfasis en la polaridad de cada dimensión mostrando una distribución de puntos orientados hacia un determinado atributo. Cada punto representa a un usuario en referencia a uno de los atributos como se muestra en la Figura 2.28. En <https://github.com/Drabin/sentiview> se puede acceder al repositorio de este desarrollo.

Un modelo diseñado inicialmente para la detección de anomalías en procesos de negocio fue desarrollado en [26]. La estructura de este modelo permitió la generalización del mismo para representar revisiones de clientes y características de producto [11], la propuesta permite encontrar correlaciones entre diferentes atributos del conjunto de datos. El desarrollo, denominado Circular Correlation Map se ilustra en la Figura 2.29. La representación muestra en el semicírculo izquierdo los atributos mencionados por los clientes en sus comentarios, donde cada uno tiene una serie de líneas que parten desde la parte central del círculo. Las líneas azules representan opiniones positivas y las rojas negativas. Una línea es dibujada cada vez que un atributo es mencionado en un comentario, así una misma revisión de un usuario puede dar origen a varias líneas en el gráfico. El semicírculo derecho muestra los ID de cada documento para poder hacer una revisión de cada opinión con la interfaz de usuario propuesta.

Algunos autores [72] integran en sus gráficos puntajes que apoyan el recurso visual, brindando un grado mayor de sensibilidad que permite observar con mayor claridad la diferencia entre las calificaciones computadas de las opiniones. La Figura 2.30 muestra un gráfico que compara las características de un dispositivo móvil mostrando las opiniones negativas con barras horizontales rojas y barras azules para representar las positivas.

2.5.4. Análisis visual orientado a la comparación de sentimientos

La detección y visualización exclusiva de sentimientos tiene también importantes campos de aplicación. Así por ejemplo se puede percibir los sen-

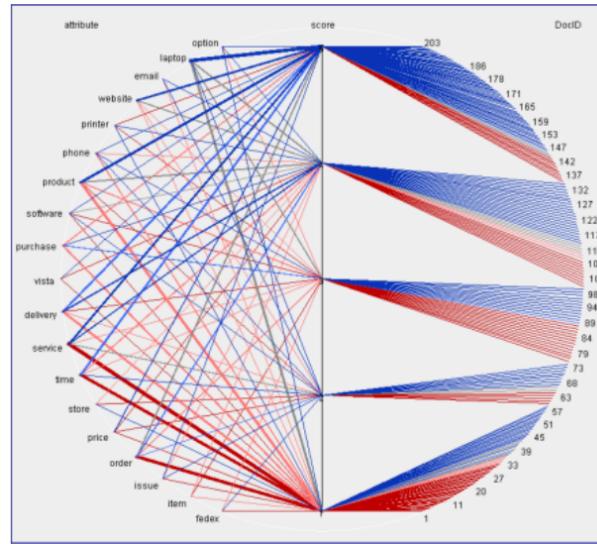


Figura 2.29: Circular Correlation Map: Comentarios de todos los clientes [11].

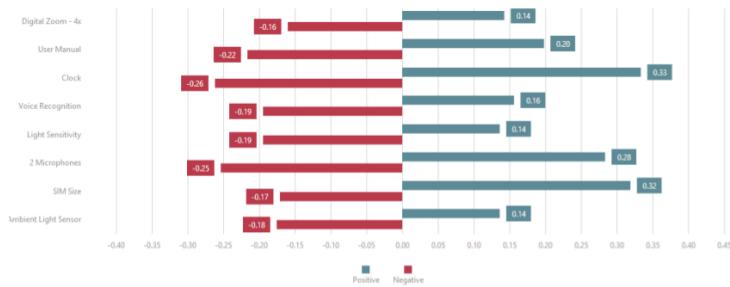


Figura 2.30: Resultados de opiniones positivas y negativas [72].

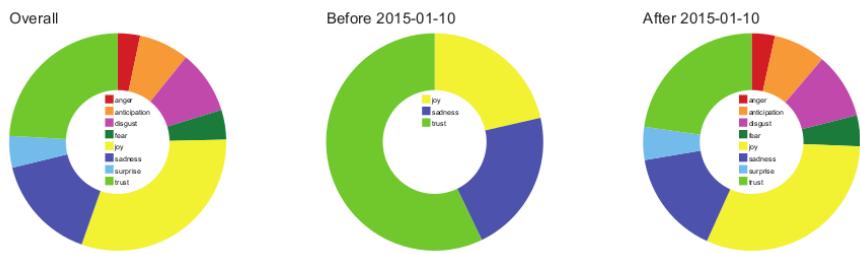


Figura 2.31: Distribución de emociones en los premios Óscar en Tweeter respecto a la palabra Boyhood [9].



Figura 2.32: Spider Char: Valor de emociones para el hashtag: AAPHelpLine [40].

timientos generados y la respuesta del público ante un estímulo, como una estrategia publicitaria. En [9] se muestra una interesante aplicación donde se estudia el comportamiento tomando medidas en diferentes intervalos de tiempo. El desarrollo sigue el comportamiento de los usuarios en Twitter de acuerdo a sus reacciones sobre los premios Óscar⁶. La visualización se da en términos de proporciones, donde se hace énfasis en los sentimientos más relevantes, como se muestra en la Figura 2.31.

El diagrama de radar es otra herramienta destacable, esta describe un patrón en función de la distribución del valor de cada uno de sus atributos. Los atributos que puede mostrar dependerá del tipo de aplicación, en [40] se utiliza para describir un patrón que representa emociones, como se observa en la Figura 2.32. El patrón obtenido se produjo con el análisis de 1000 y 10000 Tweets respectivamente y muestra una fuerte tendencia hacia el sentimiento Happiness.

Algunos autores solo hacen énfasis en el análisis con base a la proporción y cantidad de sentimientos en los documentos. En esta categorización encajan las técnicas de la sección 2.5, donde se usan típicamente diagramas

⁶Es un premio anual a la excelencia, concedido por la Academia de las Artes y las ciencias cinematográficas

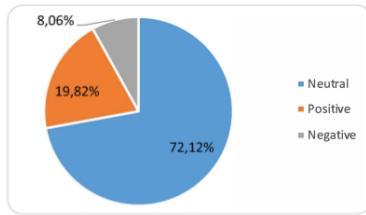


Figura 2.33: Diagrama de pie: Proporción de sentimientos positivos, negativos y neutros [20].

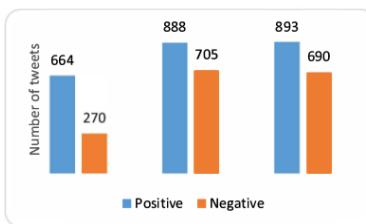


Figura 2.34: Diagrama de barras: Comparación de cantidades de sentimientos registrados en Tweets sobre tres productos [20].

de pie, histogramas y de barras en general para visualizar la proporción de opiniones en una un grupo de dimensiones. En los ejemplos de la Figura 2.2 y Figura 2.3 se visualiza la cantidad de sentimientos positivos y negativos.

Las visualizaciones enfocadas en el análisis de cantidades y proporciones pueden expresar dimensiones diferentes a las de los ejemplos dados, pudiendo así mostrar múltiples dimensiones. Un ejemplo es el caso de descubrimiento de temas en documentos donde se puede identificar la magnitud de cada campo. También se puede implementar directamente en la comparación de productos para estimar proporciones como en se desarrolló de [20]. En la Figura 2.33 se aprecia un ejemplo, allí se incorpora la dimensión neutral al modelo clásico de opiniones positivo y negativo. Otro caso se observa en la Figura 2.34, donde se muestra mediante un diagrama de barras las cantidades de Tweets obtenidos con sentimientos positivos y negativos para una serie de productos.

La cantidades pueden también usarse para conocer la distribución de documentos de acuerdo a la orientación de los mismos. Un ejemplo de esto se observa en [39], con sus representación de “Aggregation charts” que analiza la distribución de documentos de dos productos de acuerdo a los temas que tengan en común, para este caso se muestran Good, Like y Great con criterios de búsqueda Hobbit y Coca-Cola con en la Figura 2.35. Los documentos en amarillo representan los documentos seleccionados, la interfaz permite a los usuarios visualizar cada documento. El tamaño de cada documento es

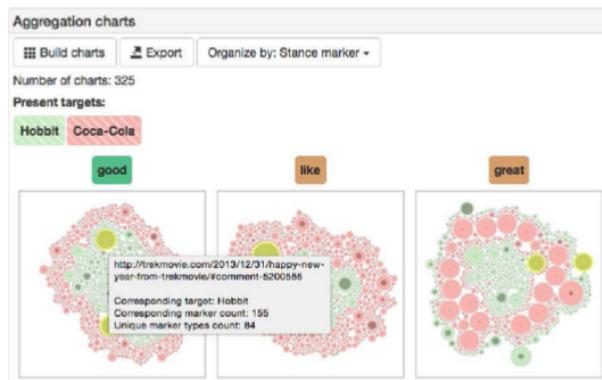


Figura 2.35: Aggregation Chart: Distribución de documentos de interés seleccionados por el usuario [39].

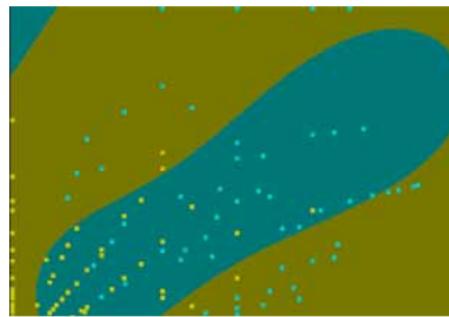


Figura 2.36: Modelo SVM para revisiones positivas: Puntos azules representan opiniones positivas y puntos amarillos negativas. El área azul identifica los límites predichos por el modelo SVM para opiniones positivas [11].

proporcional al número de instancias detectadas dentro de este. Para el caso de la Figura 2.35 la visualización esta compuesta por documentos de 1517 URLs recuperadas.

En [11] se plantea un modelo de visualización en dos dimensiones. El espacio de opiniones (x, y) depende de la cantidad de palabras y su orientación. La Figura 2.36 muestra dos zonas, donde el área de color amarillo representa una predicción de la zona de opiniones negativas realizada con SVM y el azul predice el espacio de opiniones positivas. Los puntos hacen referencia a cada opinión, el color azul se refiere a opiniones positivas y el amarillo a negativas. En la parte superior izquierda del gráfico se encontrarían opiniones negativas con mayor cantidad de términos negativos, y en la zona inferior derecha opiniones positivas con más términos positivos.

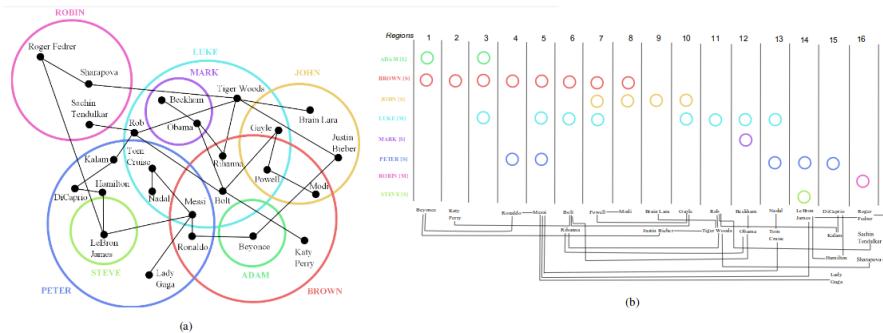


Figura 2.37: (a) Diagrama de Euler, (b) Diagrama Spherule [69].

2.5.5. Análisis visual orientado a la comparación de de usuarios e interacciones

Cuando la prioridad de visualización se enfoca en los usuarios, sus relaciones e interacciones, el modelo de visualización más aplicado corresponde al diagrama de red. El gráfico expresa en cada nodo un usuario, las interacciones o relaciones son representadas mediante aristas. La implementación de [69] con su herramienta Visual-VM es un ejemplo del uso de técnicas de agrupamiento para identificar las relaciones entre usuarios. En la Figura 2.37 se muestra un ejemplo de este tipo de representación, que emplea diagramas de Euler y Spherule para representar las conexiones. Los diagramas (a) y (b) hacen referencia a la misma representación, donde se muestran las relaciones de celebridades del espectáculo. En el gráfico se aprecian algunas relaciones indirectas entre las entidades de interés, mediante la ayuda de círculos que describen las regiones con mayor nivel de relación. La aplicación fue desarrollada en Java y se puede descargar en <https://visualvm.github.io/> [82].

En redes muy grandes, se pueden llegar a encontrar interesantes relaciones de alta complejidad. Aun así, como herramienta de visualización, la red empieza a volverse poco interpretable. Un ejemplo de lo antes mencionado se puede apreciar en la implementación de [94], en la red de Myspace⁷ [81] dentro del contexto de política mostrado en la Figura 2.38.

2.5.6. Análisis visual con enfoque geoespacial

Al análisis enfocado en la ubicación geográfica, es un problema de estudio de analistas que trabajan en áreas como el geo-marketing, estos se preocupan de analizar la localización de sus clientes, puntos e venta, etc. El objetivo con este tipo de técnicas es visualizar la información de interés sobre un mapa digital. Otra aplicación es la geo-localización de segmentos de clientes para la implementación de campañas publicitarias dirigidas. Las aplicaciones

⁷Servicio de red social digital, lanzada en agosto del 2003.

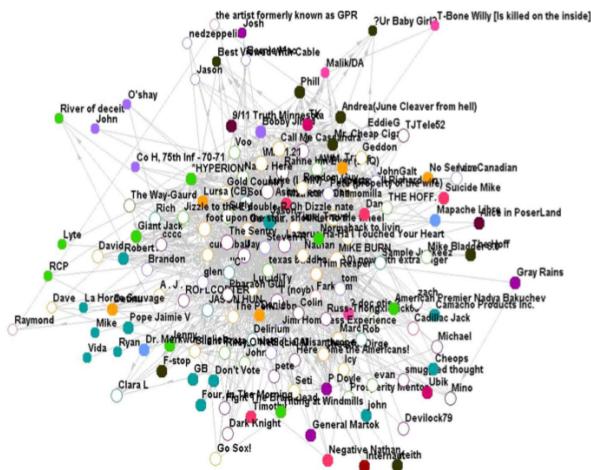


Figura 2.38: Red de MySpace en temas de política [94]: Cada nodo representa un individuo dentro de la red, las aristas muestran las relaciones de los nodos. Los individuos ubicados más cerca de la red contienen el mayor número de conexiones.

de herramientas de identificación geográfica no solo se limitan a las anteriores mencionadas, pues sirven también como apoyo en procesos estratégicos en otras áreas de aplicación como en contextos políticos, económicos y sociales.

Un ejemplo de esta aplicación es el implementado por [44], donde se desarrolló el visualizador de eventos de la Figura 2.40. Este trabajo estudia los Tweets publicados durante un temblor ocurrido en Nepal, la idea del sistema es descubrir los requerimientos de primera necesidad en cada sector. En la Figura 2.41 se puede apreciar la representación de este trabajo, allí se muestra una secuencia en el tiempo de las palabras más relevantes de cada tema: donate, tend, water, blood y medical.

Otro trabajo relevante es la aplicación “Info Vis”[68]. Mediante asociación de términos, se extraen las palabras que coinciden en simultáneas opiniones, para identificar características. Por último es tenida en cuenta la localización de opiniones en el contexto de cada característica, ejemplo de esto se aprecia en la visualización de la Figura 2.39, que muestra los sentimientos asociados a los términos “case manager” y “hawai” extraídos de opiniones con la geo-localización mostrada. El autor plantea una solución de aprendizaje no supervisado con mapas auto-organizados, al igual que [33], que plantea una visualización directa sobre la red SOM limitando la exactitud de la medida a la granularidad de las neuronas que representan la capa de salida de dicha red.

Un trabajo orientado al análisis de opiniones en el ámbito político se desarrolla en [1], [4], [13] y [75]. A través datos recolectados desde Twitter se implementa una visualización de sentimientos utilizando los emoticones



Figura 2.39: Key term Geo Map: (a) término Case Manager, (b) término Hawai [68].



Figura 2.40: Visualización de temas durante el temblor en Nepal: cada color en la figura hace referencia a alguno de los temas de tendencia, comentados en redes sociales durante el temblor; donate, tend, water, blood y medical, las concentraciones con mayor cantidad de puntos representan los temas de primera necesidad, en cada región durante la tragedia [68].

de los mensajes publicados para construir un clasificador que emplea SVM y redes de Naive Bayes.

2.5.7. Análisis visual de aspectos demográficos

La visualización de múltiples dimensiones puede llegar aumentar significativamente la complejidad de los gráficos, aun así usuarios expertos podrían llegar a sacar mayor partida de este tipo de visualización, proporcionando una visión integral y global de sus datos. Una propuesta muy interesante se desarrolla en [91] con su sistema denominado OpinionSeer. La implementación se realizó para el análisis de opiniones de un hotel, el diagrama tiene tres posibles orientaciones (positiva, negativa e incertidumbre).

OpinionSeer incluye la visualización de múltiples parámetros, entre los cuales destacan los de tipo demográfico. En la Figura 2.42 (a) se muestra la polaridad de las opiniones de acuerdo a los 6 segmentos de cliente propuestos: Business, Couples, Family, Friends getaway, None y Solo travel. Allí observa en la proyección del diagrama de barras una mayor cantidad

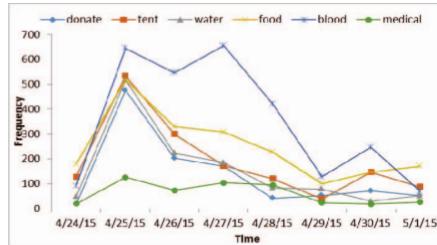


Figura 2.41: Frecuencia y distribución en el tiempo de las palabras de cada tema descubierto [68].

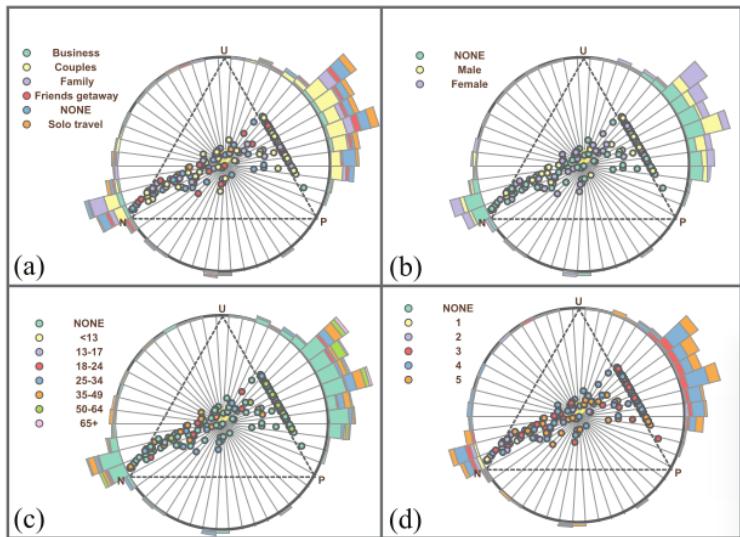


Figura 2.42: OpinionSeer: Múltiples vistas de opiniones de clientes [91].

de opiniones positivas que negativas para el segmento Couples. (b) Propone una segmentación de acuerdo al género. (c) Muestra la orientación de las opiniones de acuerdo a un rango de edad y por último (c) muestra una calificación dejada por los clientes, donde sorprendentemente se encuentran calificaciones altas en opiniones con orientación negativa.

El triángulo que describe la orientación puede ser desplazado para evitar solapamientos en el centro de la circunferencia. Se puede incluir una componente temporal discretizada, como la de la Figura 2.44 (c) y (d). Adicionalmente se pueden comparar diferentes objetos como el caso de (a) y (b) con U.S.A y China.

Otro trabajo propone un sistema denominado DemographicVis [18], que incorpora una fuente de datos desarrollada con base a una serie de encuestas digitales mediante subreddit [65]. Los datos obtenidos son empleados

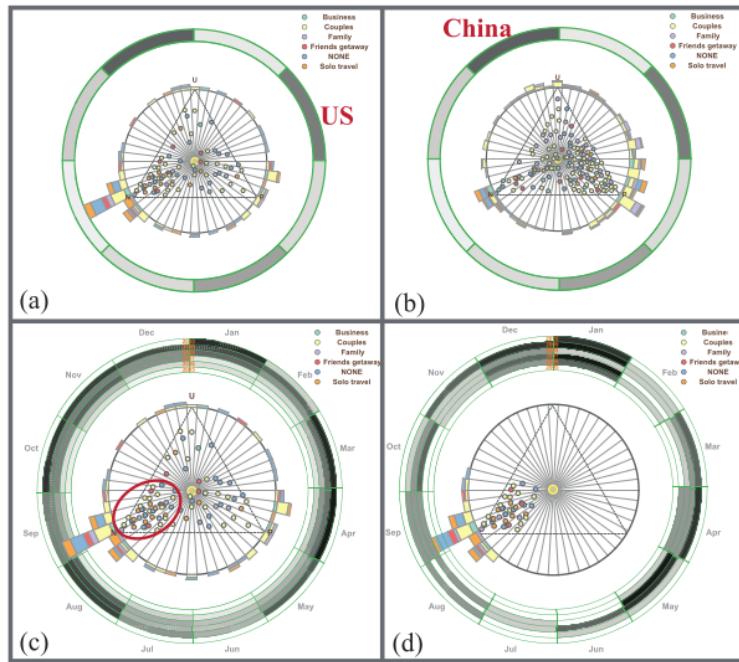


Figura 2.43: OpinionSeer: Múltiples vistas de opiniones de clientes [91].

para realizar una extracción de temas mediante la técnica Tag-LDA [45], produciendo la visualización de la Figura 2.42. Es posible visualizar las dimensiones genero, edad, edad, educación y tema, mostrando una vista de la interfaz de usuario, donde posando el puntero del ratón sobre un segmento en específico es posible ver los temas de interés que le componen. El acceso a la aplicación se encuentra en <http://demographicvis.uncc.edu/>.

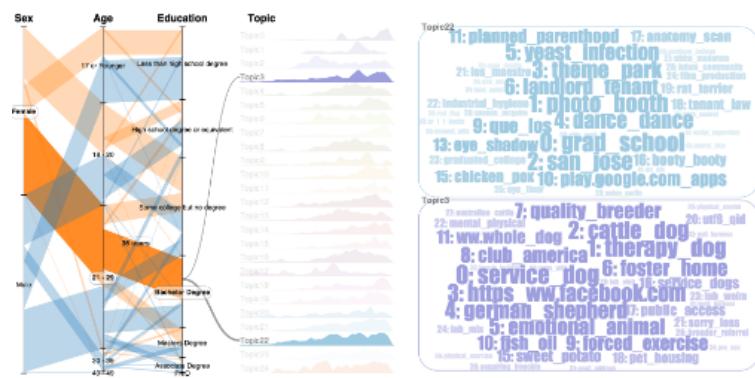


Figura 2.44: DemographicVis: Pasando el ratón sobre la interfaz en el segmento de mujer: 21-29 es posible ver los temas de interés del grupo [18].

Capítulo 3

Análisis de técnicas de visualización

Las técnicas de análisis visual adquieren relevancia desde varias dimensiones, que pueden tener origen en diversos intereses, como el tipo de problemas de negocio que resuelven (dentro del contexto de minería de opiniones para efectos de este trabajo), su facilidad de comprensión, la cantidad y el tipo de información que pueden representar. Así mismo son también de interés otras características más subjetivas que pasan a un plano más estético, como el estilo de representación y diseño intuitivo [70].

Este capítulo recoge las técnicas vistas en el capítulo 2, para realizar un análisis de las técnicas de visualización en la minería de opiniones. En la primera parte, en la sección 3.1 se exponen las principales funcionalidades de las técnicas, describiendo su potencial para representar distintos tipos de información bajo múltiples enfoques. Posteriormente, en la sección 3.2 se define el alcance de las técnicas, analizando su capacidad de representación. En la sección 3.3 se expone un análisis que incorpora una perspectiva desde un punto de vista más pragmático, considerando otros atributos más allá de las capacidades técnicas de los gráficos. La siguiente sección, la 3.4 reflejona sobre la utilizad de las técnicas, para el análisis de tópicos, características, usuarios y sentimientos. El capítulo finaliza en la sección 3.5 con unas conclusiones de los retos y oportunidades encontrados.

3.1. Principales funcionalidades:

Cada tipo de visualización exhibe ciertas propiedades y características que las hacen más, o menos adecuadas para resolver determinado tipo de problemas. Las funcionalidades principales describen el objetivo clave de visualización de cada técnica en el contexto de estudio (minería de opiniones) que se ha propuesto en la literatura.

Las técnicas estudiadas se centran en cuatro aspectos principales, que

permiten realizar análisis mediante comparaciones, datos históricos, descripciones generales, interacciones y relaciones.

3.1.1. Comparaciones:

Se aplica generalmente después de implementar técnicas para resumir la información. Así se busca encontrar diferencias contrastando aspectos particulares de los productos y clientes. El análisis visual enfocado en el producto busca por un lado evaluar las características del producto como en [43], [72] y [92]. En este tipo de visualización se contrastan las características de cada producto para distintas marcas o proveedores simultáneamente, mediante proporciones de opiniones positivas y negativas, como en las Figuras 2.25, 2.26 y 2.30, respectivamente. Estas técnicas muestran un mayor o menor nivel de granularidad y sensibilidad. Se ha observado la ponderación neta de sentimientos en cada aspecto de análisis y estas cambia con cada autor. A pesar de que en [43] y [72] se muestran más grados de libertad para representar la comparación, [92] al incluir menor sensibilidad en su análisis visual permite identificar con mayor facilidad que elemento y característica posee una mejor o peor evaluación de los usuarios.

Teniendo claro el espacio de opiniones de análisis, enfoques clásicos como el "Spider Char" [56], permiten comparar sentimientos asociados a grupos de opiniones de distinta fuente como se mostró en la Figura 2.32.

Otro enfoque clásico que se mantiene vigente es el uso de Tree Maps para comparar productos, marcas y sus características, como en la Figura 2.27 y 2.24. La cantidad de sentimientos asociados se pondera mediante escalas de color y tamaño de los bloques.

Para comparar opiniones y representar mas elementos en dos dimensiones, trabajos como [91] plantean un entorno visual con múltiples dimensiones para la comparación de tipos de clientes como en el opinion Seer de la Figura 2.7.

3.1.2. Interacciones y relaciones:

Este tipo de gráficos se concentra totalmente en el comportamiento de los usuarios. El análisis se puede enfocar en las interacciones en un hilo de opiniones como en el desarrollo de [12], que se puede observar en la Figura 2.22. Otro enfoque en la misma vía es el planteado por [69], de la Figura 2.37 representado mediante un diagrama de red, que según el tipo de análisis, puede llevar a una topología particular de la red. En [11] se plantea un esquema que relaciona las interacciones en base a sentimientos positivos y negativos de un mismo grupo de usuarios en el tiempo como se vio en la Figura 2.19.

Dentro del contexto de relaciones en [84], se desarrolla un modelo que representa la polaridad de distintas opiniones, conectándolas cuando aparecen

usuarios con intereses comunes como en la Figura 2.6.

3.1.3. Comportamiento histórico (temporal):

El comportamiento histórico en el contexto de análisis visual de opiniones busca incorporar tres dimensiones en la visualización; el tema de opinión, los sentimientos asociados (típicamente positivos y negativos) y el tiempo que puede ser representado con distintos niveles de granularidad. [11] como se mostró en la Figura 2.20 propone dos histogramas, uno para sentimientos positivos y otro para negativos que muestra los temas tratados mediante bloques de palabras. Este enfoque puede ser adecuado en contextos con cantidades de palabras limitadas (por ejemplo donde se tienen claros previamente los temas de interés), debido a que comentarios de usuarios con mucha diversidad de temáticas pueden saturar este tipo de gráfico, haciéndolo difícil de interpretar. Otro elemento que no es tenido en cuenta, es la cantidad de palabras de un mismo tema, pues este tipo de visualización no distingue la frecuencia.

Un modelo alterno para visualizar de manera más intuitiva la cantidad de opiniones sobre un tema y sus sentimientos asociados en el tiempo es propuesta por [90]. Mediante una eficiente representación de áreas de color, se muestra cada sentimiento asociado a un tema en el tiempo, como en la Figura 2.21. Al igual que el modelo anterior este tipo de gráfico se puede llegar a saturar con mucha diversidad de temáticas en un mismo intervalo de tiempo.

Un tercer trabajo relevante [44] incluye un componente que no fue tenido en cuenta en los trabajos anteriores, este muestra la frecuencia de temáticas de interés en el tiempo. La Figura 2.41 expone con claridad el comportamiento en el tiempo de la popularidad de dicho tema. Este enfoque podría implementar una mejora razonable con un proceso de análisis de sentimientos, para agregar valor en el contexto de minería de opiniones. Se podría representar mediante líneas independientes del mismo tema los sentimientos asociados a cada tema.

Un último enfoque [17], ilustrado en la Figura 2.5, muestra el histórico de opiniones resumidas donde cada opinión es ponderada con un valor numérico que representa una polaridad (positiva ó negativa). Este enfoque brinda un grado mayor de granularidad en la evaluación de sentimientos de las opiniones, permitiendo conocer la tendencia global del espacio de opiniones en el tiempo. El problema de este tipo de gráfico es que debido a su sencillez, se desconoce el periodo de tiempo de las opiniones y los temas en los que se centran.

3.1.4. Descripción y caracterización de distribuciones (Resumen):

Algunas técnicas de análisis visual no se centran en el detalle, estas pretenden mostrar aspectos más generales que permitan exponer una visión más general sobre las opiniones y documentos de análisis. Este tipo de análisis visual es usado para representar: Cantidad, proporciones, distribuciones y frecuencias de opiniones, sentimientos, usuarios y tópicos de interés. Dentro de este grupo entran todas las técnicas de visualización clásicas como el diagrama de pie (Figuras 2.3 y 2.31), diagrama de barras (Figura 2.2), los tree maps (Figura 2.24 y 2.27) y el diagrama de radar de la Figura 2.32.

Además de las anteriores mencionadas, destacan algunas técnicas como [39], por incorporar una novedosa forma de ilustrar la distribución de documentos u opiniones por temática. Un ejemplo fue la visualización vista en la Figura 2.35, que representa con zonas de distinto color y tamaño los tópicos de interés.

Otro trabajo relevante es [84], que realiza una modificación de un histograma convencional, dividiendo las barras en dos colores para representar así la proporción de opiniones sobre un tema en un periodo establecido. La Figura 2.18 y el astrolabio de la Figura 2.6 propuesto por [17] son un claro ejemplo de este tipo de visualización. El objetivo central de estas técnicas, se centra en la distribución de las opiniones y su orientación en determinadas temáticas.

3.1.5. Descripción de parámetros demográficos:

Mediante cookies o formularios web, es posible añadir más dimensiones a las capas de visualización. Estas dimensiones permiten visualizar cada uno de los segmentos de clientes y su comportamiento, mediante la representación gráfica de variables demográficas. Los trabajos que incluyen este tipo de representación son pocos. El Opinion Seer [91] y DemographicVis [18] plantean dos geometrías muy interesantes. El primero, se enfoca en el análisis de características de productos/servicios. El segundo, se centra en el análisis de tópicos o temáticas de interés. Este tipo de análisis visual tiene un gran campo de aplicación en marketing. Es capaz de representar múltiples variables en un mismo gráfico, lo que permite comparar varias dimensiones simultáneamente. Estas visualizaciones muestran un panorama global del problema. Una desventaja que podrían tener, es la complejidad misma de los gráficos, que agregaría al grafo una baja capacidad de interpretación, en especial para usuarios no expertos.

3.1.6. Representación geográfica (espacial):

Los datos de geo-localización pueden venir de tres clases de dispositivos: ordenadores, terminales móviles y tabletas. Mediante geo-localización

IP y sistemas de GPS del propio terminal, se puede conocer con una buena sensibilidad la ubicación de un usuario, bastará con que el individuo tenga habilitado su dispositivo para la navegación GPS y cuente con acceso a Internet. Este tipo de visualizaciones permite relacionar sentimientos asociados a opiniones, con regiones del espacio. Los trabajos encontrados muestran distintos grados de granularidad, representando países, ciudades o incluso partes específicas dentro de una ciudad ó al interior de una gran superficie, como un campus universitario. Se busca con este tipo de visualización la clasificación de sentimientos y tópicos en áreas del espacio.

3.2. Capacidad de las técnicas

Las técnicas de visualización en minería de opiniones buscan representar gráficamente información relevante, resumiendo a distintos niveles de abstracción la información disponible. Dependiendo del tipo de técnica estas empiezan a mostrar limitaciones en su capacidad de representación, principalmente en tres aspectos; en el número de dimensiones, la cantidad de datos y el tipo de datos que pueden representar.

3.2.1. Tipo de datos que pueden procesar:

La información fuente para la minería de opiniones se recupera en formatos semi estructurado y no estructurado. La información procesada incluye texto, que según la funcionalidad de las técnicas de visualización, pueden definir el tipo de datos textuales de entrada requeridos. De los documentos de texto se extraen temáticas relevantes, sujetos de interés, características de productos/servicios y sentimientos asociados. Como se vio en la Sección 3.2, la estructura de datos puede incluir adicionalmente meta datos, que incorporan información correspondiente a puntuaciones numéricas, tiempo, características demográficas (ubicación, edad, genero etc.) y polaridad de las opiniones.

A manera de ejemplo, para implementar el gráfico de la Figura 4.13, es requerido un proceso de extracción de temas relevantes que deben de ir acompañados de meta datos de tiempo, para situar cada palabra clave en un espacio de tiempo determinado. Otros tipo de visualización mas complejos pueden requerir meta datos relevantes a la ubicación, genero y edad de los usuarios, para ubicar su opinión en un segmento de clientes predeterminado, como es el caso del OpinionSeer [91] visto en la Figura 2.7. Alguno tipos de visualización tienen requerimientos de entrada mas simples, ya que solo cuentan la cantidad de opiniones y los sentimientos asociados para representar una vista general en un diagrama. Un claro ejemplo es el diagrama de pie, como en la Figura 2.33. La Figura 2.40 muestra otro ejemplo de aplicación y uso de meta datos, para este caso una extracción de temas relevantes

en el texto es acompañada de una visualización en un mapa, gracias a los meta datos de localización geográfica: latitud y longitud.

Los meta datos pueden ser suministrados por la fuente, pero en determinados casos se puede inferir del texto, algunos de ellos mediante técnicas de extracción de datos latentes, que se sale del alcance de este trabajo.

3.2.2. Cantidad de datos que pueden representar:

Grandes volúmenes de datos puede dar origen a gráficos poco interpretables. En este aspecto las técnicas de resumen de datos juegan un papel fundamental disminuyendo la granularidad del problema. La dificultad en mención se materializa principalmente cuando el número de entidades a representar aumenta cuando la cantidad de datos es superior. Una entidad puede ser una opinión, grupo de opiniones, personas o cualquier otro tipo de instancia que represente el gráfico.

Aunque las técnicas de resumen de datos pueden ayudar a solucionar ciertos problemas, algunos tipos de análisis visual por su naturaleza pueden estar predispuestos a mostrar mejor o peor grandes cantidades de datos. Un ejemplo típico de un gráfico afectado por la cantidad de datos son los que involucren mapas, como el visto en la Figura 2.40. Otro ejemplo es el circular correlation Map [11] visto en la Figura 2.29, que realiza una serie de trazas correspondientes a los atributos del producto mencionados por los clientes, con sus respectivos sentimientos asociados. Una gran cantidad de trazas podría saturar la visualización hasta un punto que estas no se pueden reconocer con claridad.

Algunos tipos de análisis visual se ven afectados por su geometría, como es el caso del Opinion Ring [19] y Opinion Seer [91], que debido a su forma circular puede llegar acumular zonas de alta densidad de puntos en el centro para algunos casos.

Los tipos de visualización que representar interacciones como los grafos de red vistos en la Figura 2.38 o las interacciones en hitos de opinión de la Figura 2.6, se saturan fácilmente con apenas unos cientos de datos, haciendo difícil representar grandes cantidades de interacciones (de usuarios en comentarios etc.).

3.2.3. Número de dimensiones que pueden representar:

Este tipo de restricción es casi un estándar en todas las técnicas estudiadas. Los elementos a representar tienen generalmente 2 ó 3 dimensiones, cada dimensión corresponde a alguno de los tipo de datos que pueden procesar vistos anteriormente.

Un par de técnicas de análisis visual destacan mostrando mas de 3 dimensiones, en un mismo gráfico. Este es el caso del Opinion Seer [18] y [91] de la Figura 2.44, que incluyen una gran diversidad de parámetros demográficos.

cos, acompañado de las dimensiones clásicas para visualizar la polaridad de sentimientos, temas y atributos de interés.

3.3. Características subjetivas

Algunas características de los métodos de análisis visual en la minería de opiniones adquieren importancia en un plano subjetivo, donde la experiencia y preferencias de los usuarios pueden llegar a afectar significativamente en la percepción que tienen sobre la representación. Existen algunos estudios cuya evaluación se centra en medir el tipo de características mencionadas. El estudio descrito a continuación se desarrolló mediante un grupo de evaluadores constituidos por una población que incluye personas expertas y no expertas en las temáticas de estudio. El trabajo en mención es [70], allí se referencia 10 métricas que definen las bondades de las técnicas de visualización en 6 áreas, como se menciona a continuación.

Cada área de evaluación responde las siguientes preguntas de las técnicas de visualización:

1. **Visual impact:** ¿Es agradable a la vista de los usuarios?
2. **Overall performance:** ¿Es fácil de entender? , ¿Es amigable con el usuario?
3. **Overall desing style:** ¿Es informativa? , ¿Es intuitiva?
4. **Information quality:** ¿Es útil?, ¿Es comprensible?
5. **Visual representation model:** ¿Es buena la comparación de datos?, ¿Es buena el estilo de representación de datos?
6. **Information presentation model:** ¿Se requiere conocimiento previo para entender la visualización?

Los autores evaluaron cuatro tipos de visualizaciones:

1. **Hierarchical:** donde se incluyo el Tree map [23] y Visual summary [56].
2. **Radial:** donde se tuvieron en cuenta las técnicas Opinion wheel [91] y Rose plot variation [25].
3. **Graph:** donde estudiaron el Coordinated graph [11], Positioning map [52], Line graph and pie chart [49] y Comparative relation map [92].
4. **Bar chart:** donde incluyeron Glowing, Bar chart with symbols [86] y Bar chart [43].

Del estudio realizado destacó con mayor relevancia la técnica Glowing bars, por su fácil comprensión y diseño informativo, seguida de las siguientes técnicas: El tree map con una métrica amigable al usuario destacable. Coordinated graph y Comparative relation map revelaron un estilo de representación relevante a sus evaluadores. Rose plot destacó por su potencial utilidad. Visual summary se mostró agradable a la vista y amigable al usuario y Bar chart with symbols destaco por su alta capacidad de comprensión.

3.4. Utilidad de técnicas de análisis visual

Las técnicas vistas se muestran más adecuadas para resolver determinados tipos de problemas. En esta sección se realizará un resumen enfocado en ilustrar el objetivo principal de cada técnica. Se encontraron cuatro objetivos de análisis: basado en temas o tópicos, en características o atributos, centrado en usuarios y enfocado en sentimientos. Dentro de cada uno de los objetivos de análisis, y según lo visto previamente en este capítulo, se encontraron seis dimensiones principales de análisis. Las dimensiones pueden ser desde un contexto espacial (geográficos), temporal (históricos), de interacción, comparación, resúmenes y demográfico.

Objetivos de análisis: En la literatura se encontraron 4 enfoques principales en lo que análisis visual de opiniones concierne. Estos se diferencian principalmente en el pre-procesamiento de los textos para enfocar el análisis visual en el área de interés. Posterior al pre-procesamiento se sigue un proceso muy similar en cada uno de los enfoques. Se realiza una evaluación de los sentimientos, un resumen de la información y posteriormente se aplica una técnicas de visualización en alguna de las 6 dimensiones especificadas. A pesar de que el último enfoque hace parte del flujo de los anteriores (sentimientos), se especifican por separado como caso especial en vista que bajo esta perspectiva se realiza una evaluación enfocada exclusivamente en los sentimientos encontrados. Las tablas 3.1 y 3.2 resumen varios de los trabajos encontrados bajo esta perspectiva.

1. **Temas o tópicos:** Se caracteriza por una etapa de pre-procesamiento dedicada a la extracción de temas relevantes en el texto, para comprender la trama principal de las opiniones. Algunos autores como [84], aplican técnicas de PLN para descubrir dichos tópicos en el texto. Otros definen previamente los temas de estudio, haciendo un esfuerzo en identificar elementos relacionados exclusivamente a los temas de pre-establecidos.
2. **Características o atributos:** Esta etapa requiere previamente una etapa de filtrado para identificar opiniones. Una vez se han identificado, mediante técnicas de PLN se realiza la extracción de las características u atributos de cada opinión

Enfoque/Dimensión	Espacial	Temporal	Interacción
Temas o Tópicos	[44]	[11] [84] [11] [90]	[90] [12]
Carac. o atributos	[68],[33]	[84]	[84]
Usuarios			[94] [90] [69]
Sentimientos	[13],[68],[75],[4]		[11]

Tabla 3.1: Enfoques de técnicas de visualización parte 1: según el objetivo de visualización

Enfoque/Dimensión	Comparación	Resumen	Demográfico
Temas o Tópicos	[19],[73]	[39]	[18]
Carac. o atributos	[43] [92] [56] [72]		
Usuarios			[91]
Sentimientos	[17]	[9] [20] [40]	[91] [18]

Tabla 3.2: Enfoques de técnicas de visualización parte 2: según el objetivo de visualización

3. **Usuarios:** El análisis centrado en el usuario se enfoca principalmente en estudiar las interacciones de usuarios en una red. Este tipo de interacciones se puede dar de múltiples formas, un caso típico es el seguimiento de opiniones de usuario en un grupo.
4. **Sentimientos:** Este enfoque busca identificar y evaluar los principales sentimientos en el grupo de documentos dado.

La tabla 3.4 y 3.3 muestran un análisis desde la perspectiva algorítmica, clasificando los trabajos de la tabla 3.1 y 3.2 en 3 grupos; desarrollos que han empleado exclusivamente aprendizaje supervisado o no supervisado y enfoques mixtos. Las técnicas de aprendizaje automático y PLN en general son utilizadas tanto para identificar patrones de interés en los documentos (como se vio arriba) como para construir modelos para la evaluación de sentimientos. Bajo el enfoque supervisado son ampliamente utilizados clasificadores con Naive Bayes y SVM. Bajo el enfoque no supervisado se encuentran técnicas de agrupamiento, reglas de asociación, DLA y enfoques PLN basados en diccionario en todas sus formas. Los enfoques mixtos hacen referencia a la combinación de los dos anteriores.

La selección de la técnica de análisis visual a utilizar dependerá de tres cuestiones principales; el objetivo de visualización, la dimensión que se desee proyectar y el tipo de datos del que se disponga.

Enfoque/Dimensión	Espacial	Temporal	Interacción
Supervisado	[33],[68],[13],[4], [75]		
No supervisado	[44], [68]	[11], [84], [84], [90]	[84],[94], [11], [90]
Mixto			[12], [69]

Tabla 3.3: Enfoques de técnicas de visualización parte 1: según el tipo de técnicas utilizado

Enfoque/Dimensión	Comparación	Resumen	Demográfico
Supervisado	[19]	[9]	[91]
No supervisado	[17], [56], [72]	[40], [39]	
Mixto	[23], [43], [92]	[40]	[18]

Tabla 3.4: Enfoques de técnicas de visualización parte 2: según el tipo de técnicas utilizado

3.5. Retos y oportunidades

En este capítulo se expuso las principales propiedades, funcionalidades y características de las técnicas de análisis visual. Las tablas 3.1 y 3.2 muestran una gran cantidad de trabajos que tratan la dimensión de comparaciones con extracción de características y extracción de tópicos con análisis temporal. Aún así la tabla de referencia muestra que no están cubiertas con rigurosidad muchas regiones. Algunas presentan un verdadero reto, como el análisis de usuarios y sus aspectos demográficos ó el análisis de atributos del producto en el contexto espacial, ya que la complejidad de las variables es de alta dimensionalidad y las técnicas de análisis visual no pueden representarlas con facilidad. En algunos casos como los mencionados arriba, aparece una dificultad latente al tratar de combinar distintos enfoques y dimensiones, como se observa claramente en los campos vacíos de las tablas 3.1 y 3.2 . Aun así otros campos solo están vacíos por aparente falta de interés. Por citar un ejemplo, es bien conocido que existe una cantidad relevante de trabajos de análisis espacial y sentimientos, que no son extendidos al análisis de opiniones, como en [75], donde se realiza un análisis de sentimientos dentro de un campus universitario. Muchas de las técnicas en realidad se han implementado en contextos similares pero estos no han sido generalizados de manera amplia en la minería de opiniones para el análisis de productos.

El análisis de grandes volúmenes de datos trae importantes retos para las técnicas de análisis visual, las técnicas vistas en las sección anterior muestran unas visualizaciones limitadas en cuanto a capacidad de representar múltiples dimensiones, esta dificultad se acentúa aun más cuando se busca simultáneamente representar grupos de productos. La cantidad de dimensiones a representar (características ó tópicos principalmente) y el número de instancias (clientes o productos por ejemplo) están claramente limitadas. Ha

pesar de los distintos esfuerzos de varios autores, que han experimentando con diversas geometrías, colores, matices, texturas etc. No se ha podido extender exitosamente el alcance de las dimensiones, para grandes volúmenes de datos. La inclusión de un gran número de entidades trae con sigo una saturación irremediable de los gráficos, haciendo estos incomprensibles.

Dentro del contexto de análisis visual en minería de opiniones de productos, hay un número de técnicas con alto potencial pendientes por implementar, destacando entre estas el diagramas de burbujas, que además permiten una representación de hasta de 4 dimensiones, para una cantidad de instancias moderadamente alta. Otro candidato con grandes posibilidades es el diagrama de Venn, que ofrece la posibilidad de contrastar las similitudes de productos ó clientes en un gráfico de dos dimensiones. Este último destaca la intersección de sus instancias (entiéndase por instancia productos o clientes).

Cada técnica de visualización revisada en el estado del arte se construyó orientada a representar una dimensión concreta, aun así estás permiten su generalización para representar otro tipo de información, a manera de ejemplo las técnicas mostradas de análisis de tópicos son en buena medida compatibles con la visualización de características y viceversa

Las técnicas ilustradas en el capítulo 2, pueden ser aprovechadas en mejor medida con técnicas de pre-procesamiento que resuman la información. Se podrán configurar también algunas pequeñas modificaciones que brinden más funcionalidades, como el uso del color, variación del tamaño y formas de los objetos de análisis. Así mismo el uso de los semi planos restantes, que en la mayoría de gráficos están en desuso, podrían permitir la representación de más instancias.

Capítulo 4

Librería de visualización de opiniones

En este capítulo se realiza una descripción de la estructura del paquete de software desarrollado. En la Sección 4.1 se expone la estructura del paquete: métodos, clases y atributos. En la misma sección se ilustra la relación de las clases, estructura de datos y parámetros de entrada y salida de toda la librería. La clase de pre procesamiento del paquete se expone en la Sección 4.1.1. En las Secciones 4.1.2 y 4.1.3 se trata el desarrollo para implementar extracción de tópicos y características respectivamente. Este capítulo finaliza con la Sección 4.1.4, donde se describe todo lo relacionado a la estructura de las clases para la generación de gráficos, con algunos ejemplos de implementación.

4.1. Librería en Python de análisis visual

La librería propuesta ha sido desarrollada en Python, por tres razones principales: es un lenguaje compatible con programación orientada a objetos, es uno de los lenguajes de programación con mayor comunidad de desarrolladores en cuestión de ciencia de datos, y por último, contiene librerías de pre procesamiento y visualización orientadas a objetos compatibles con el alcance planteado para este trabajo.

La librería desarrollada [21], contiene cinco clases: la primera llamada "Preprocesamiento" se encarga del procesado para transformar conjuntos de datos en estructuras compatibles con las librerías de visualización de Python. Las cuatro clases restantes, llamadas "graphicsGeneralSummary", "graphicsComppare", "graphicsSummary" y "graphicsTime", contienen los métodos necesarios para generar los gráficos. Las dependencias de la librería incluyen los paquetes de Python: gzip, pandas, matplotlib, numpy y squarify. Las clases se relacionan como se muestra en la Figura 4.1, Las clases "graphicsGeneralSummary", "graphicsComppare", "graphicsSummary" heredan los

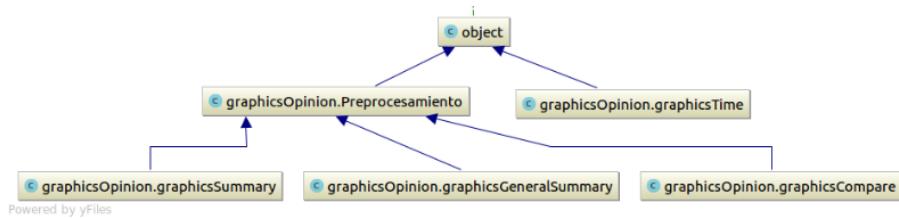


Figura 4.1: Esquema de clases de la librería de análisis visual desarrollada.

métodos y atributos de la clase padre "Preprocesamiento" la clase "graphicsTime" no tiene ninguna relación con las anteriores, esta contiene sus propios sus métodos de pre procesamiento y visualización.

El diagrama de clases se muestra en detalle en la Figura 4.2, donde las clases se señalan en azul con la letra "c", los métodos en rojo con la letra "m" y los atributos en amarillo con la letra "f". Las cinco clases contienen en total 22 métodos, de los cuales 10 desarrollan tareas de pre procesamiento. Los 12 restantes generan gráficos. Un total de 14 atributos contiene la librería, con dos tipos de estructura de datos: "arrays" de tipo "numpy" y "data Drame" de tipo "pandas".

La librería es de código abierto ¹ y está disponible en Git Hub en el siguiente enlace :

<https://github.com/jespinosal/visuaReviewsAnalysis>

La librería esta desarrollada en un único fichero de Python llamado "graphicsOpinion.py", este contiene las clases tanto de pre procesamiento, como construcción de gráficos. En el método principal, en el "main" del paquete se describen varios flujos con ejemplos que muestran como utilizar la librería propuesta, ilustrando el flujo a seguir para instanciar cada Clase y modificar los atributos de la misma cuando corresponda.

4.1.1. Pre procesamiento de datos

A continuación se exponen los métodos de procesamiento implementados más relevantes. Cada técnica buscará transformar la fuente de datos en formas compatibles, con cada técnica de visualización propuesta.

- **summarizeProductTopics:** Tiene como objetivo agrupar los tópicos y/o características de cada producto, resumiendo así la calificación promedio y número de opiniones de cada valoración del producto. Para lograrlos utiliza funciones de agregación SQL. Esta función devuelve un nuevo conjunto de datos donde se retorna para cada producto la cantidad de opiniones por tópico ó característica encontrada.

¹OSI, "Open source initiative". 1998, <https://opensource.org/docs/osd>.

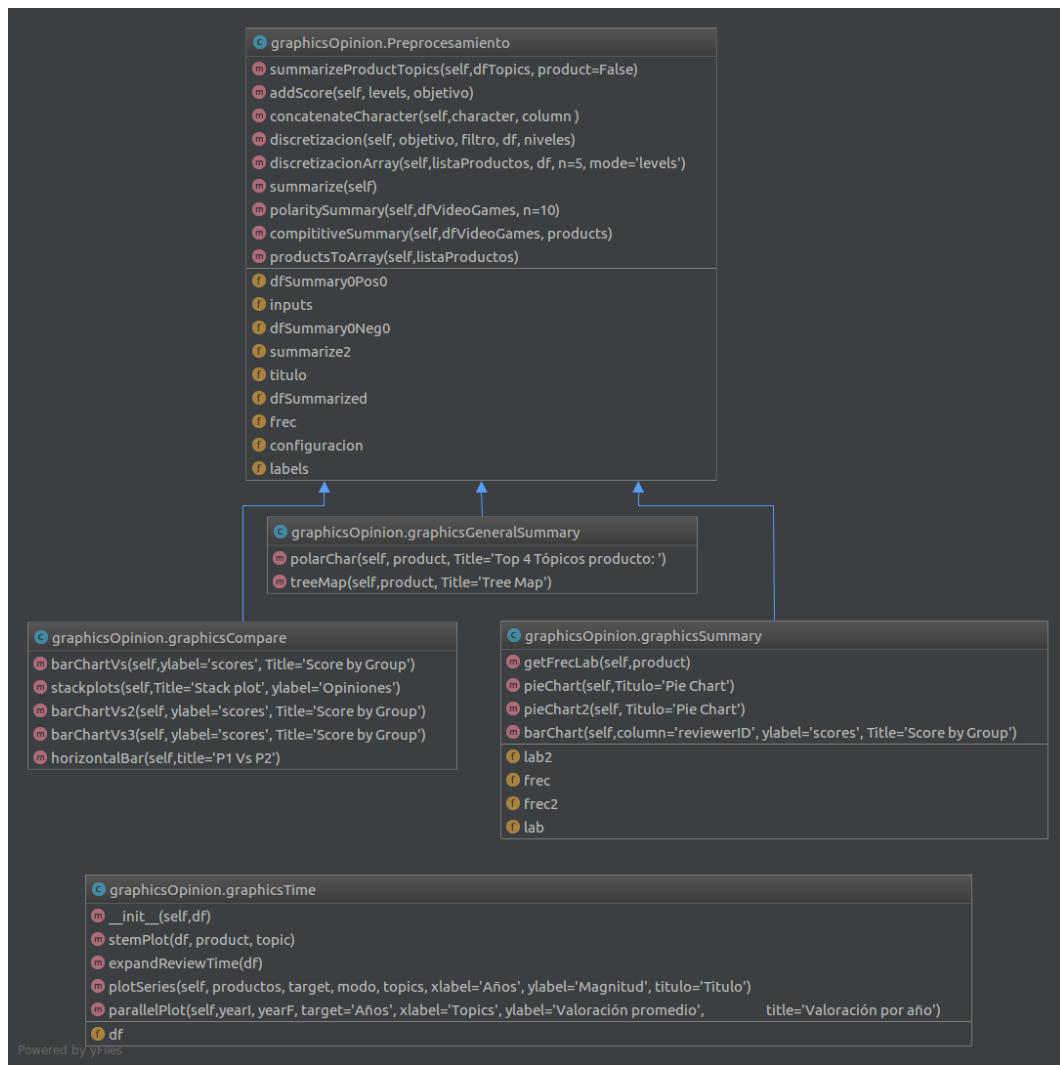


Figura 4.2: Diagrama detallado de clases de la librería de análisis visual desarrollada.

- **AddScore:** Su objetivo es resumir la información, este método discretiza variables numéricas reales, transformándolas en variables en categóricas. Con esto se logra una representación más simple que permita resumir y visualizar una cantidad finita de estados por cada opinión. Este método define varios niveles discretos para remplazar dichas etiquetas por su valor real más similar. Así una variable real con un rango entre 0 y 5, podría transformarse en una categórica de 3 niveles: negativo, neutral y positivo, o 5 niveles: muy malo, malo, regular, bueno y excelente. Esta función se usa en concreto para discretizar el parámetro “overall” que contiene la valoración del producto del cliente.
- **discretizacionArray:** Devuelve el valor discretizado de la variable real “overall”, en forma de array. El objetivo es poder automatizar las parámetros de entrada de algunos gráficos de librerías como matplotlib². Esta función tiene como parámetro adicional una función de selección, para filtrar un grupo de productos determinados, produciendo así un arreglo de salida que incluye solo los productos deseados. Este proceso ofrece la posibilidad de realizar un análisis particular de subgrupos, para comparar productos.
- **discretizacion:** Realiza el remplazo de variables numéricas a categóricas a bajo nivel, y es llamado dentro de los métodos “AddScore” y “discretizacionArray”. Dispone de tres parámetros de configuración: objetivo (columna sobre la cual se realizará el cálculo), filtro (Id del producto a procesar), df (conjunto de datos) y niveles (nivel de granularidad de la transformación de número a categoría, 3 ó 5 niveles). Si se especifica un valor distinto de 3 ó 5 el método retorna las etiquetas originales de la variable objetivo.
- **summarize:** Permite crear un conjunto de datos que contiene todas las combinaciones de tópicos con productos, para que mediante operaciones de agregación e imputación se consiga imputar cada producto con valores de tópicos vacíos cuando corresponda.
- **polaritySummary:** Haciendo uso del parámetro “overall” directamente como una variable que expresa el nivel de satisfacción del cliente con el producto, se procede a procesar esta variable para asignar un valor a los sentimientos generados por la experiencia de usuario. Se dividen las respuestas en sentimientos con tendencias positivas y negativas. Para dicho fin se desarrolló la función “polaritySummary”, que permiten cambiar la escala de “overall” de 0 a 5 a una escala de -2.5 a 2.5. La función devuelve al final una medida del promedio de las opinio-

²Biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación Python

nes multiplicada por un peso, que para efectos de este trabajo es la cantidad de opiniones del tópico ó característica de cada producto.

- `competitiveSummary`: Para poder comparar parejas de productos en un espacio cartesiano, se requiere generar una distribución que ocupe los semi planos del mapa con los valores correspondientes. Así a manera de ejemplo, si se quiere utilizar un diagrama de barras que representa un producto en el eje positivo del plano vertical, se puede utilizar el plano negativo vertical para poder representar otro producto y comparar. Para esto la función “`competitiveSummary`” ofrece una alternativa, ya que distribuye la polaridad de los valores a visualizar para poder representar dos productos en un mismo plano cartesiano. Esto permite que se pueda usar tanto el eje vertical como horizontal para representar dos visualizaciones de productos simultáneamente, reflejados en uno de los ejes.
- `expandReviewTime`: Las visualizaciones de series temporales requieren para su representación consideraciones especiales en lo que a formato de tiempo respecta. Para poder implementar este tipo de representaciones se requiere de una etapa de pre procesamiento en dos fases, la primera crea la serie temporal con las variables de interés. La segunda implementa operaciones de agregación en la ventana temporal dada, para resumir la información. Todo esto se hace mediante la función “`expandReviewTime`”, que entre otras cosas da formato a los meta datos de tiempo para definir intervalos. Para poder gestionar múltiples ventanas de tiempo, “`expandReviewTime`” da niveles de granularidad menores al campo original, pudiendo así extraer solo años, meses o días.
- La función `ProductsToArray` genera un arreglo dado un grupo de productos y dos conjuntos de datos, uno de opiniones positivas y otro de opiniones negativas, generado previamente con el método “`polaritySummary`”. La salida es un arreglo que contiene un resumen de las valoraciones de cada tópico, por cada producto en cada una de las polaridades.

Los atributos de las clase pre procesamiento se describen a continuación:

- `dfSummary0Pos0`: Este atributo es modificado por el método “`polaritySummary`” de la clase “`Preprocesamiento`”. Contiene una muestra del conjunto de datos original con las opiniones positivas son filtradas, resumiendo la cantidad de opiniones y la polarización de las mismas en un nuevo campo denominado “`summary`”.
- `dfSummary0Neg0`: Realiza la misma acción de “`dfSummary0Pos0`” sobre las opiniones negativas.

- summarize2: Es modificado por método “summarize” de la clase “Pre-procesamiento” y realiza una operación de agregación para imputar los productos que no dispongan de tópicos o características de estudio.
- freq: Es afectado por el método de “discretizacion” de la clase “Pre-procesamiento” y guarda una lista con el resumen de opiniones, resumiendo cantidad de opiniones o valoraciones de las mismas, según se indique en el método “discretizacion” en su parámetro de entrada “objetivo” (“overall” para valoraciones y “summary” para cantidad de opiniones).
- labels: Lista de etiquetas categóricas a representar, según el caso pueden ser productos, tópicos o características. Los atributos “freq” y “labels” conforman la pareja de tuplas necesarias para mapear las coordenadas (variable,valor) en un plano de dos dimensiones.
- titulo: Es una cadena de caracteres de texto que representan el título del gráfico a representar.
- configuracion: Es un diccionario que contiene un resumen de las configuraciones aplicadas al objeto instanciado. Así si un objeto tiene una configuración determinada, “configuracion” recoge dicha información. La información almacenada es utilizada para configurar los títulos de los gráficos, mostrando así los filtros aplicados y productos de análisis principalmente.

4.1.2. Extracción de tópicos

Las técnicas de extracción de tópicos buscan encontrar un grupo de temas entorno a unas palabras en un conjunto de documentos. LDA ha sido seleccionado para este trabajo, el algoritmo se referencia en la literatura como uno de los más representativos para dicho fin. La implementación se realizará mediante Python, apoyado en la librería NLTK para realizar todo el pre procesamiento de texto y tareas de procesamiento del lenguaje natural necesarias. La implementación de LDA se realizará a través de la librería “gensim” [66] de Python.

El alcance de la extracción de tópicos para efectos de este trabajo consiste en agregar un nuevo campo en el conjunto de datos que asigne el tópico más probable a cada opinión. La metodología para la extracción de tópicos se sigue casi en su totalidad la mencionada en la Sección 2.2.1, implementando los pasos a continuación:

- Analizador léxico (Tokenizer)
- Normalización del texto
- Eliminación palabras vacías(Stop words)

- Filtro Porter Stemmer
- Construcción del corpus
- Filtrado tokens baja frecuencia
- Construcción de matriz de frecuencias de palabras
- Construcción del modelo LDA
- Análisis de tópicos encontrados.
- Etiquetado de opiniones

Inicialmente se implementa un analizador léxico que divide las frases en tokens de palabras. Cada palabra es normalizada en letras minúsculas, se eliminan símbolos y palabras vacías. Para aumentar la frecuencia de las palabras en los documentos, se implementa una transformación de palabras a su forma raíz, mediante el algoritmo de Porter. Posteriormente se filtran todas las palabras con frecuencia de aparición baja, con la que posteriormente se construye una matriz de frecuencias para construir el modelo LDA.

La configuración inicial del modelo ha sido ajustada para identificar 10 tópicos. Se aplica el modelo al mismo grupo de observaciones utilizado para el entrenamiento. A continuación se realiza un etiquetado del tópico de mayor probabilidad para cada opinión. Por último se incorpora al conjunto de datos fuente, un nuevo campo denominado “máximos” que contiene la etiqueta del tópico más probable.

4.1.3. Extracción de características

Inicialmente se hace un etiquetado de todos las palabras de los documentos, para identificar la estructura de la palabra. A continuación se eliminan todas las cifras numéricas (para poder aumentar la frecuencia de aparición de grupos de palabras). Palabras implícitas y explícitas que representen características deben de ser reemplazadas por una forma estándar, así ejemplo característica implícitas como “Mb” ó “Gb” tienen el mismo significado que “almacenamiento”, todo esto para referirse al mismo aspecto, la “Memoria” del dispositivo (esto beneficia también el aumento de la frecuencia de aparición de cada característica). Convertir el grupo de documentos a N-grams, teniendo en cuenta que varios trabajos importantes 3-gram ha mostrado buenos resultados. Una vez realizado el POS tag aparecerán algunas secuencias con POS tag duplicados, estos se deberán de enumerar en la misma secuencia para poder procesarlo más adelante. Buscando aumentar aún más la frecuencia de aparición de las palabras se deberá de implementar el algoritmo de Porter para transformar cada elemento en su forma raíz. Por último se procede a guardar las secuencias de N-grams en un data set transaccional.

El proceso descrito arriba permitirá la aplicación de reglas de asociación, con la intención de identificar reglas que permitan la extracción de características, entendiendo por esto como la extracción de opiniones de las singularidades de un determinado producto, como la capacidad de memoria de almacenamiento para un terminal móvil, la cantidad de mega píxeles para una cámara fotográfica, o la duración de la batería de una laptop a manera de ejemplo.

Las reglas de asociación según [43] mostraron un buen resultado filtrando reglas cuya estructura contiene sustantivos y verbos en el antecedente y características en el consecuente. Con la intención de poder usar las reglas encontradas como herramientas para la extracción de características, se hace necesario la transformación de estos en patrones de lenguaje. La idea es identificar características candidatas en nuevos documentos de opiniones con los patrones definidos, en un proceso iterativo donde se resolvieran paso a paso conflictos, se agregarán nuevos sinónimos de características y se depurarán frases poco relevantes al contexto.

Para efectos de este trabajo se propone una metodología de extracción de características rápida con base a lo planteado por [43], donde se toman varias elementos para aprovechar la estructura de datos del conjunto de datos de Amazon. El enfoque tiene 3 etapas:

- Análisis: Selección productos sobre meta datos
- Filtrado: Productos sobre meta datos
- Análisis: Selección productos sobre opiniones
- Filtrado: Productos sobre opiniones
- Análisis: Selección características
- Filtrado: Características

El objetivo de este proceso es identificar todos los sustantivos en el conjunto de datos, para reconocer el universo de palabras que hacen referencia a los productos de interés y sus potenciales características a analizar. Para esto se construyeron las funciones “mainNounsExtraction” y “mainNouns”, la primera que identifica productos y la segunda sus características.

Análisis y selección productos sobre meta datos

La función “frequencyAnalysis” devuelve las frecuencias de strings anidadas en listas y arreglos. El trabajo de análisis inicial consiste en identificar manualmente las frecuencias de cada palabra en los campos que contiene nombres de productos para definir las posibles representaciones del nombre de un producto y sus atributos.

Filtrado de productos sobre meta datos

Conocidos los nombres de los productos, se etiqueta cada opinión con su producto respectivo creando un nuevo campo en el conjunto de datos para dicho fin. A continuación se filtra cada producto creando un conjunto de datos independiente por cada producto a analizar.

Análisis y selección productos sobre opiniones

Mediante el método “mainNounsExtraction” se realiza un análisis de cada producto, para las principales formas en las que se hace mención del mismo. Las distintas formas se guardan en un diccionario de sinónimos para su posterior uso.

Filtrado de productos sobre opiniones

Con el diccionario de términos que se referencia a cada producto se utiliza la función “iterateFilter”, que realiza un filtrado del conjunto de datos para extraer opiniones que contiene sustantivos indicados en su parámetro de entrada.

Análisis y selección características

La función “mainNouns” extrae la frecuencia de sustantivos del conjunto de datos de cada producto. Para cada comentario se hace un etiquetado gramatical de las palabras de cada opinión, para filtrar los sustantivos presentes en cada una que representen características. Las que resulten relevantes deberán de incluirse a el diccionario de sinónimos.

Filtrado de características

Nuevamente se hace uso de la función “iterateFilter” para filtrar los comentarios que contengas las características seleccionadas.

4.1.4. Técnicas de visualización

La implementación que se describe a continuación, desarrolla una librería de análisis visual, donde se incorporan las técnicas más relevantes vistas en el estado del arte, utilizando librerías de visualización como: matplotlib [31] y paquetes de calculo científico como numpy [83] y de análisis de datos como pandas [47].

Descripción de las clases

A continuación se describe la estructura de métodos y atributos de las clases que generan las gráficos en la librería:

Clase “graphicsGeneralSummary”:

■ Métodos:

- polarChar: Genera una visualización de radar, sus parámetros de entrada son dos, una lista de categorías a representar y otra lista de los valores numéricos que representan cada categoría.
- treeMap: Este método devuelve un gráfico de tipo “treemap”, los parámetros de entradas de esta función son dos, una lista de las categorías a representar y una lista con los valores numéricos asociados a cada categoría.

Clase “graphicsCompare”:

■ Métodos:

- barChartVs: Genera una secuencia de barras verticales que representan cada categoría dentro de un producto (características o tópicos), los parámetros de entrada incluyen las variables internas de la clase “inputs” y “labels” que son arreglos donde cada columna representa una variable categórica y cada fila un producto con sus correspondientes valores numéricos, para representar valoraciones promedio o cantidad de opiniones.
- stackplots: Esta visualización representa cantidades categóricas, como pueden ser tópicos o características. El gráfico se genera con las variables internas de la clase “inputs” y “labels”. Por defecto la visualización se hace sobre el conjunto completo de datos, pero con el parámetro de entrada “products” es posible filtrar una lista de productos definida.
- barchrVs2: Devuelve una visualización que compara enésimos productos, entre sus parámetros de entrada se incluye “products”, la lista de productos a evaluar. Con la lista de productos se generan las variables internas de la clase “inputs” y “labels” (necesarias para indicar las coordenadas en el gráfico), donde cada producto es representado en las filas de estos arrays, que contienen las categorías a analizar y sus valores respectivos.
- barChartVs3: Compara dos productos con barras en un mismo eje vertical, los productos son indicados en su parámetro de entrada “products”. Sobre los productos indicados se generan dos arrays en las variables internas de la clase “inputs” y “labels”.
- horizontalBar: Compara dos productos con barras en un mismo eje horizontal, los productos son indicados en su parámetro de entrada “products”. Sobre los productos indicados se generan dos arrays en las variables internas de la clase “inputs” y “labels”.

Clase “graphicsSummary”:

■ Métodos

- `getFrecLab`: Abstira de la clase padre “Preprocesamiento” el atributo “dfSummarized”, que es filtrado por los atributos de la clase “graphicsSummary” en “lab”, “frec”, “lab2” y “frec2”
- `pieChart`: Genera un gráfico de pie con las variables internas de la clase “frec” y “lab”. Por defecto el gráfico se genera sobre todo el conjunto de datos, pero es posible filtrar por productos señalando en el atributo “products” los productos a tener en cuenta.
- `pieChart2`: Muestra un gráfico de pie con paleta de colores para representar una dimensión más, tiene como parámetros de entrada a “freq”, “lab” y “lab2”.
- `barChart`: Genera un diagrama de barras de las categorías a analizar, con los valores del array “frec2”, donde se puede especificar que se desea visualizar, valoraciones ó cantidad de opiniones.

■ Atributos:

- `lab`: Lista de categorías a evaluar, pueden ser tópicos o características.
- `frec`: Esta lista contiene la cantidad de opiniones encontradas de cada categoría de estudio en “lab”.
- `lab2`: Es una lista que almacena para cada categoría de “lab” su valoración categórica de la opinión (mala, buena, etc.)
- `fec2`: Agrupa en un array las listas: “lab”, “frec”, “lab2” y la valoración promedio por cada categoría a evaluar

Clase “graphicsTime”:

■ Métodos

- `init`: Método de inicialización de parámetros que guarda en el atributo “df” en el conjunto de datos de estudio.
- `expandReview`: Procesa “df” para estructurar los meta datos de tiempo en el conjunto de datos. Los datos de tiempo con el formato: MMddyy H:mm:ss zzz. Dicha estructura se formatea en sus sub categorías en un nuevo campo del conjunto de datos por año, mes y día.
- `stemPlot`: Este método genera un gráfico tipo “stem” cuyo parámetro de entrada es lista de productos a tener en cuenta.

- **plotSeries:** Muestra la serie temporal del conjunto de datos “df” respecto a alguno de sus parámetros numéricos, como pueden ser cantidad de opiniones, valoraciones promedio y relevancia de la opinión. También se puede visualizar resultados de variables categóricas obtenidas por extracción de tópicos o características. Los parámetros de entrada son: “productos” (lista de productos a tener en cuenta), “target” (variable de “df” a representar en la serie temporal), “topicos” (lista de tópicos a tener en cuenta en la visualización).
- **parallelPlot:** Construye una visualización de ejes verticales que representan como cambia una variable categórica respecto a una numérica en el tiempo. Entre sus parámetros de entrada están: “yearI” (un entero que indica el tiempo de inicio del análisis), “yearF” (un entero que indica el tiempo de fin del análisis) y “target” (define la granularidad del análisis, día, mes o año)
- **Atributos:**
 - **Atributos**
 - **df:** Conjunto de datos de estudio.

Pie Char

Este tipo de visualización requiere un proceso previo de discretización de los sentimientos asociados a las evaluaciones de los usuarios a un determinado producto. Para preprocesar la información se hace uso de una librería de pre procesamiento propia PreprocessingLibrary.py, donde mediante el método “discretizacion” se resumen los sentimientos asociados en 5 o 3 categorías.

Las librerías de análisis visual incluyen dos métodos, pieChar y pieChar2, donde la segunda incluye las funcionalidades de la primera añadiendo el uso de etiquetas adicionales para representar la dimensión color.

El gráfico final muestra la proporción de sentimientos asociados según el nivel de granularidad seleccionado, como se muestra en la Figura 4.3. Los gráficos de diagrama de pie de las figuras 4.3 y 4.5, son generados mediante las siguientes líneas de comando:

```

1 graficosResumen = graphicsSummary()
2 graficosResumen.summarizeProductTopics(dfVideoGames)
3 graficosResumen.addScore(5, 'overall')
4 graficosResumen.concatenateCharacter('T', 'maximos')
5 graficosResumen.getFrecLab('0700099867')
6 graficosResumen.pieChart(Titulo='Pie Chart')
7 graficosResumen.pieChart2(Titulo='Pie Chart')

```

Este tipo de gráfico resulta útil para visualizar variables con un número limitado de etiquetas. La Figura 4.4 muestra una representación de los

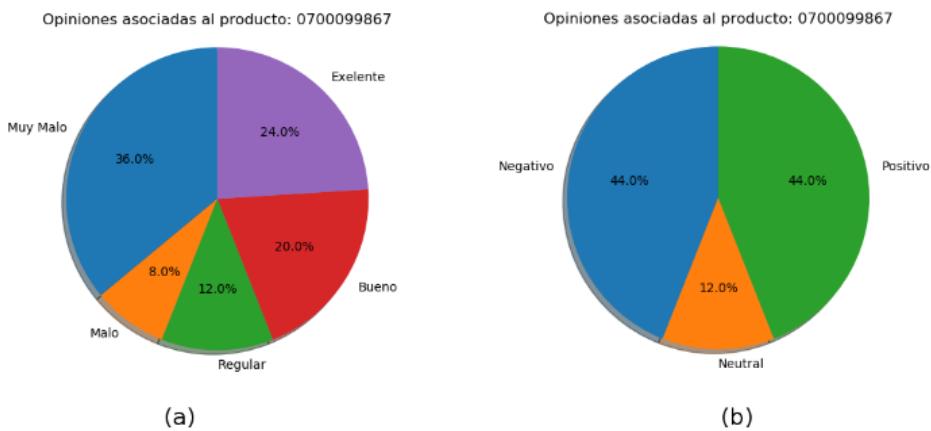


Figura 4.3: Diagrama pie: (a) Distribución 5 categorías, (b) Distribución 3 categorías.

tópicos asociados al producto de referencia, donde se muestra en (a) una representación de la cantidad de opiniones y en (b) el valor promedio de la valoración del producto.

Mediante el apoyo de etiquetas de color se representa simultáneamente una evaluación de sentimientos por tópico del producto de referencia como se muestra en la Figura 4.5.

Bar Char

Al igual que el diagrama de pie, el diagrama de barras es útil para representar variables categóricas, haciendo necesario la implementación de funciones para transformar variables numéricas en un número finito de categorías.

Los métodos que permiten representar el diagrama de barras en la librería son barChart y barChart2, el segundo incluye las mismas funcionalidades del primero con la excepción que este admite el procesamiento de array de datos, para comparar múltiples productos en una misma visualización. Para el pre procesamiento se utilizan tres métodos, discretization, discretizacionArray y summarize, las dos primeras resumen productos y grupos de productos respectivamente y la segunda devuelve estadísticas brutas de todo el conjunto de datos.

La visualización de la Figura 4.6 b muestra una representación de una valoración de clientes de un producto en 5 categorías y la Figura 4.6 a muestra una de 3 categorías para el mismo producto. Las líneas de comando para generar la visualización de la Figura 4.6 se enseñan a continuación:

```
1 | graficosResumen = graphicsSummary()
```

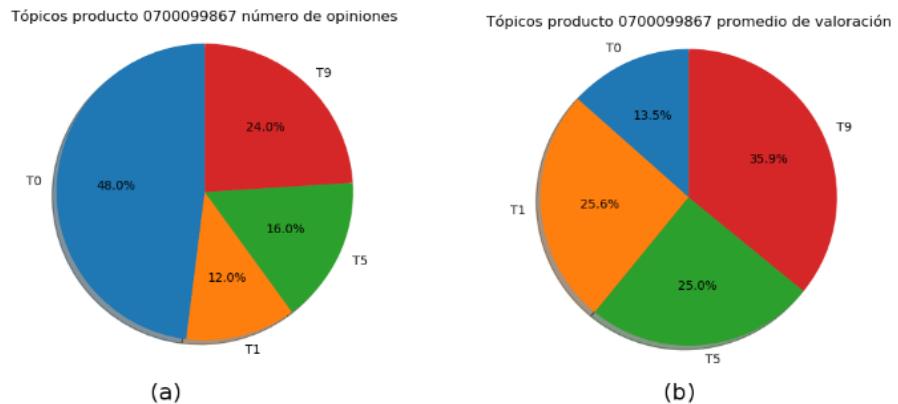


Figura 4.4: Diagrama pie: (a) Proporción cantidad de opiniones, (b) Proporción promedio valoración.

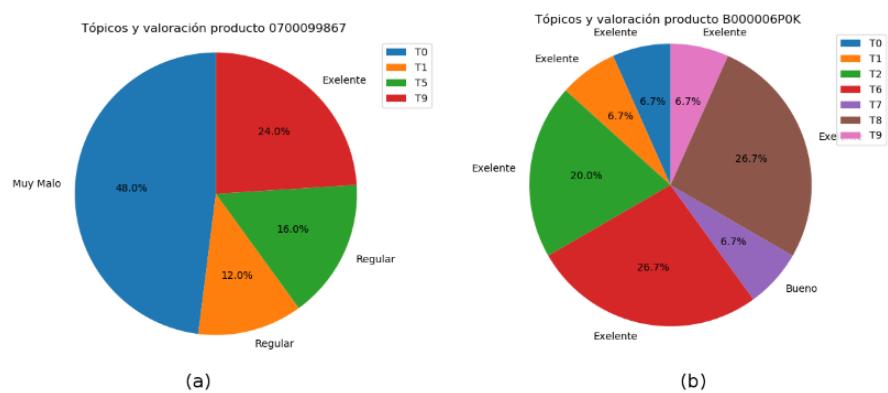


Figura 4.5: Diagrama pie: (a) Proporción Tópicos y valoración producto 0700099867, (b) Proporción Tópicos y valoración producto B000006P0K.

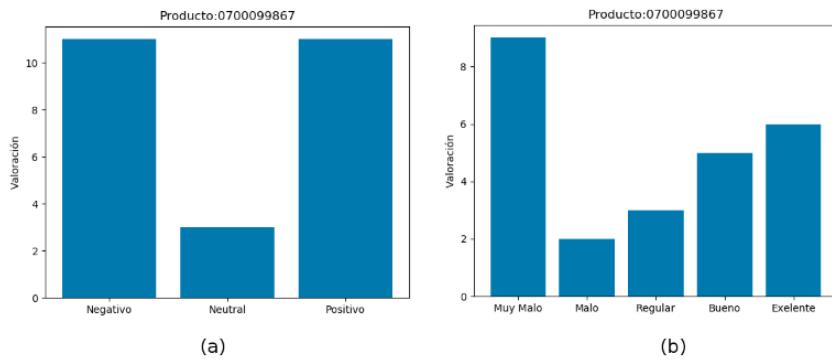


Figura 4.6: Diagrama barras: Valoraciones de producto 0700099867 (a) 3 categorías (b) 5 categorías.

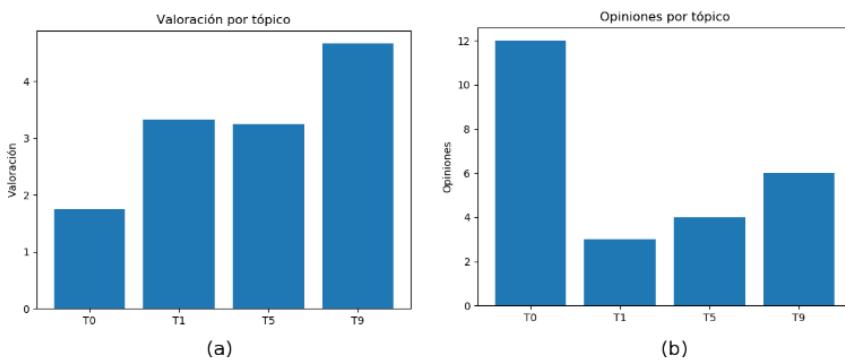


Figura 4.7: Diagrama barras: Análisis de tópicos (a) Promedio de valoraciones (b) Cantidad de opiniones.

```

2 | graficosResumen.summarizeProductTopics(dfVideoGames)
3 | graficosResumen.addScore(3, 'overall')
4 | graficosResumen.concatenateCharacter('T', 'maximos')
5 | graficosResumen.getFrecLab('0700099867')
6 | graficosResumen.barChart('overall', 'Valoracion', 'Valoracion por topico')
  )
```

Haciendo foco en los tópicos descubiertos en el conjunto de opiniones se plantea las visualizaciones de la Figura 4.7, donde se muestra un resumen de los tópicos encontrados para el producto 0700099867, haciendo énfasis en el promedio de valoraciones en *a* y en la cantidad de opiniones en *b*.

La propuesta de la Figura 4.8 consiste en una representación múltiple donde se muestra la cantidad de opiniones con respectiva etiqueta de valoración. La misma figura compara simultáneamente 3 productos en 5 dimensiones, definidas por el conjunto de 5 etiquetas: Muy malo, malo, regular,

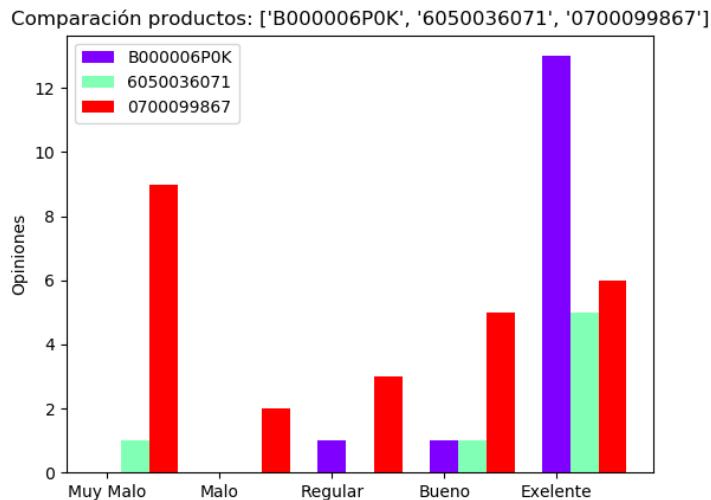


Figura 4.8: Diagrama barras: Comparación de valoraciones y cantidad de opiniones de productos.

bueno y excelente. En la visualización destaca el producto B000006POK (en lila) que tiene el mayor número de opiniones con valoraciones positivas en la categoría excelente, duplicando en número a los otros dos productos mostrados.

Las representaciones que se pueden implementar dependerá en gran medida de las dimensiones que contenga el conjunto de datos, lo que vendrá a continuación es un proceso de operaciones de agregación e imputación de datos para obtener los arreglos de datos en un formato propio para la representación. La Figura 4.9 muestra una representación de valoraciones de tres productos en la dimensión de 10 tópicos.

Polar Char

Representaciones radiales pueden brindan una visualización de alta simetría, para comparar magnitudes. En la Figura 4.10 se observa el producto de referencia en 4 distintas dimensiones, que representan cuatro tópicos distintos. El mismo grafo permite identificar la magnitud de cada opinión, representando el número de opiniones con su radio. El ancho de cada representación de los tópicos representa la valoración de las opiniones, siendo las más ancha la opinión más positiva y la mas angosta la más negativa. Para esta visualización se representa un tópico cada noventa grados, pero en caso de haber más tópicos el espacio se verá dividido en más porciones. La Figura 4.10 se genera mediante las siguientes líneas de comando:

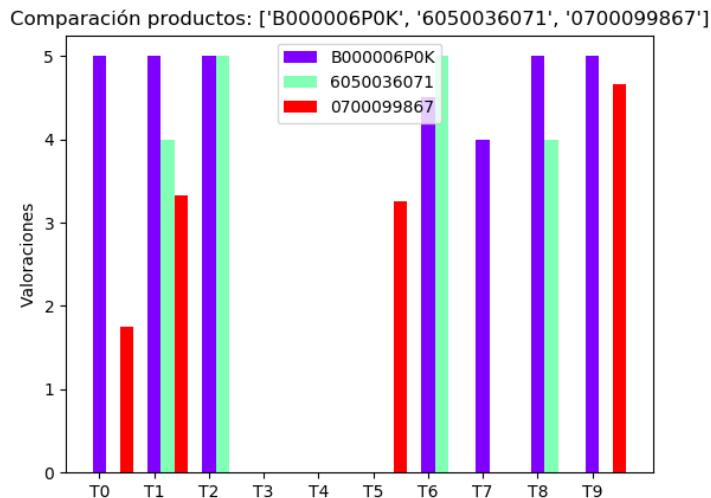


Figura 4.9: Diagrama barras: Comparación de valoraciones producto por tópico.

```

1 graficoResumenGeneral=graphicsGeneralSummary()
2 graficoResumenGeneral.summarizeProductTopics(dfVideoGames)
3 graficoResumenGeneral.polarChar(product='0700099867', Title= 'Score by
    Topic')

```

Tree Map

Como se observó en el capítulo 2, el Tree Map por su simplicidad se postula como uno las técnicas de visualización más fáciles de interpretar. Su implementación se hace sobre variables categóricas, por lo que deberá de implementar un proceso de pre procesamiento para las variables numéricas, haciendo uso de las funciones summarizeProductTopics, para resumir la información y AddScore para convertir los datos a un formato de categorías.

La visualización de la Figura 4.11 muestra una representación de las valoraciones de un producto, que contiene opiniones con calificaciones que encajan en 4 de las 5 etiquetas disponibles. La Figura 4.11 es generada con las siguientes líneas de comando:

```

1 graficoResumenGeneral=graphicsGeneralSummary()
2 graficoResumenGeneral.summarizeProductTopics(dfVideoGames)
3 graficoResumenGeneral.addScore(5, 'overall')
4 graficoResumenGeneral.concatenateCharacter('T', 'maximos' )
5 graficoResumenGeneral.treeMap(product='0700099867', Title= 'Score by
    Topic')

```

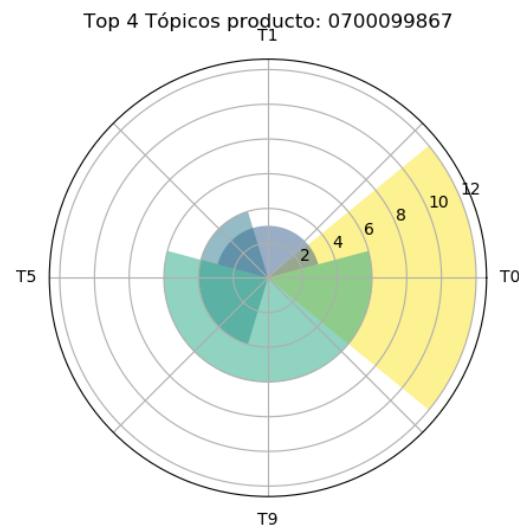


Figura 4.10: Diagrama polar: Comparación de valoraciones producto.



Figura 4.11: Diagrama de mapa de árbol: Resumen de valoraciones de producto.

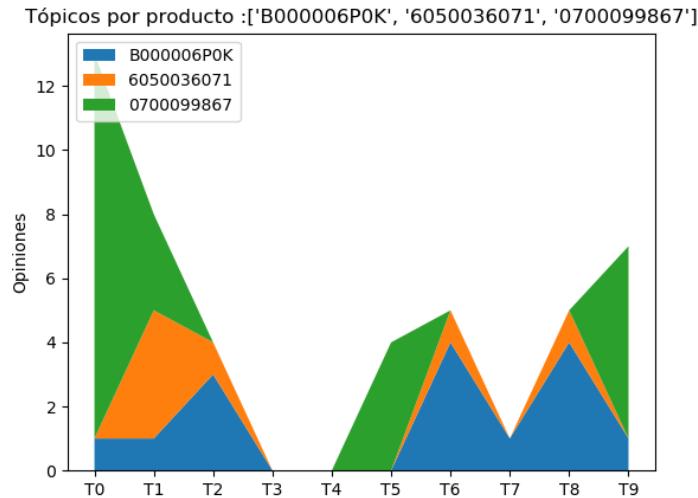


Figura 4.12: Stack plot: Cantidad de opiniones por tópico de tres productos.

Stack plot

El área como representación de magnitudes muestra interesantes facultades para comparar elementos con categorías comunes. La Figura 4.12 representa la cantidad de opiniones de tres distintos productos en una gama de 10 tópicos distintos, donde el producto en verde muestra una predominancia en los primeros 4 tópicos y el azul domina en los últimos 5. El gráfico de la Figura 4.12 se genera con las siguientes instrucciones de código:

```

1  graficosComparacion = graphicsCompare()
2  graficosComparacion.summarizeProductTopics(dfVideoGames)
3  graficosComparacion.addScore(5, 'overall')
4  graficosComparacion.concatenateCharacter('T', 'maximos')
5  graficosComparacion.summarize()
6  graficosComparacion.discretizacionArray(['B000006POK
    ','6050036071','0700099867'],graficosComparacion.summarize2,5,
    mode='summary')
7  graficosComparacion.stackplots>Title='Stack plot', ylabel='Opiniones
    ')

```

Series temporales

Las valoraciones realizadas por los clientes y los tópicos de cada opinión están acotados dentro de una ventana temporal, haciendo posible incluir esta dimensión en la visualización. Mediante las funciones expandReviewTime y plotSeries se pre procesan las opiniones para generar series temporales de variables de interés.

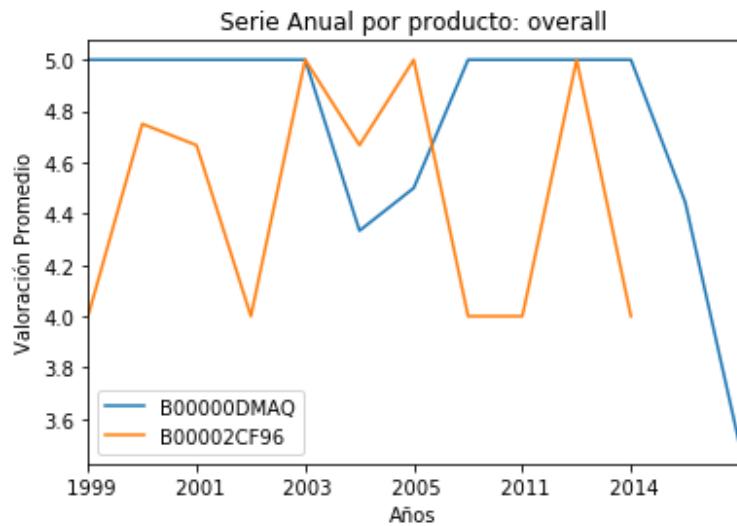


Figura 4.13: Series temporales: Valoración promedio de opiniones por año.

La 4.13 representa una visualización del promedio de opiniones de dos productos en un periodo de 16 años, donde en la mayoría del periodo las opiniones mantienen una calificación alta. Las gráficas de las Figuras, 4.13, 4.14 y 4.15, son generadas con las siguientes líneas de comando:

```

1 graficosTiempo= graphicsTime(dfVideoGames)
2 graficosTiempo.parallelPlot(2007,2014,target = 'Anos', xlabel = ,
   Topics', ylabel = 'Valoracion promedio', title = 'Valoracion por
   año')
3 graficosTiempo= graphicsTime(dfVideoGames)
4 graficosTiempo.plotSeries(productos = ["B00000DMAQ","B00002CF96"],
   target ='overall',modo = "asin", topics=[], ylabel= 'Valoracion
   Promedio', xlabel = 'Anos')
5 graficosTiempo.plotSeries(productos = ["B00000DMAQ","B00002CF96"],
   target ='summary',modo = "asin", topics=[], ylabel= 'Opiniones',
   xlabel = 'Anos')
6 graficosTiempo.plotSeries(productos = [],target ='overall',modo =
   maximos", topics=[0,1,2,3], ylabel= 'Valoracion Promedio', xlabel =
   'Anos')
7 graficosTiempo.plotSeries(productos = [],target ='summary',modo =
   maximos", topics=[0,1,2,3], ylabel= 'Opiniones', xlabel = 'Anos')

```

El promedio no tiene mucho valor sin conocer la cantidad de opiniones que respaldan dicha calificación, la Figura 4.14 compara los productos de referencia, mostrando un pico sobre el final del periodo del producto representado en azul, respaldando el valor promedio de 9 opiniones con una calificación promedio de 4.6.

Los tópicos asociados a cada opinión pueden también representarse. La Figura 4.15 muestra una visualización del histórico del promedio de valora-

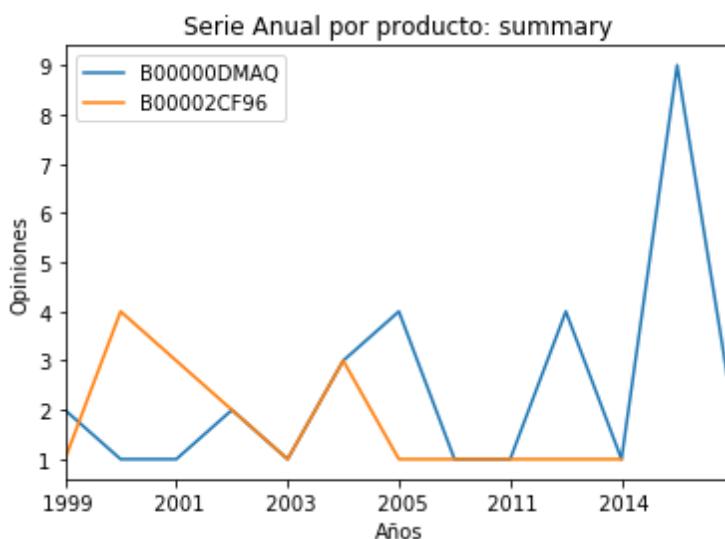


Figura 4.14: Series temporales: Cantidad de opiniones por año.

ciones de las opiniones para un grupo de tópicos seleccionados.

El histórico de la cantidad de opiniones asociado a cada tópico se puede visualizar en la Figura 4.16 mostrando un aumento considerable de la cantidad de opiniones a partir del año 2008 para el tópico 0 y el tópico 1.

Parallel plot

La visualización de variables categóricas adquiere relevancia cuando la visualización de opiniones hace énfasis en variables de este tipo, como lo pueden ser los tópicos y características de productos. Este tipo de visualización muestra un panorama de como es el flujo de opiniones de cada categoría en una dimensión particular. Un ejemplo de esto se ilustra en la Figura 4.17, donde se visualiza la valoración promedio de opiniones por tópico de cada año. En la Figura 4.17 a se calculó sobre un histórico de 16 años y en la Figura 4.17 b se realizó sobre 5 años. Esta visualización puede llegar a saturarse en gran medida cuando el número de categorías en el eje horizontal es muy grande o el número de elementos a visualizar es ligeramente alto. Esto se puede observar en las Figuras citadas anteriormente.

La visualización del parallel plot requiere una representación tabular gestionada mediante el método `pivot_table` de la librería Pandas. Simultáneamente es necesario definir la granularidad de la variable temporal mediante el método `expandReviewTime` disponible en el librería de pre procesamiento `PreprocessingLibrary.py`. El código que genera la Figura 4.17 se muestra a continuación:

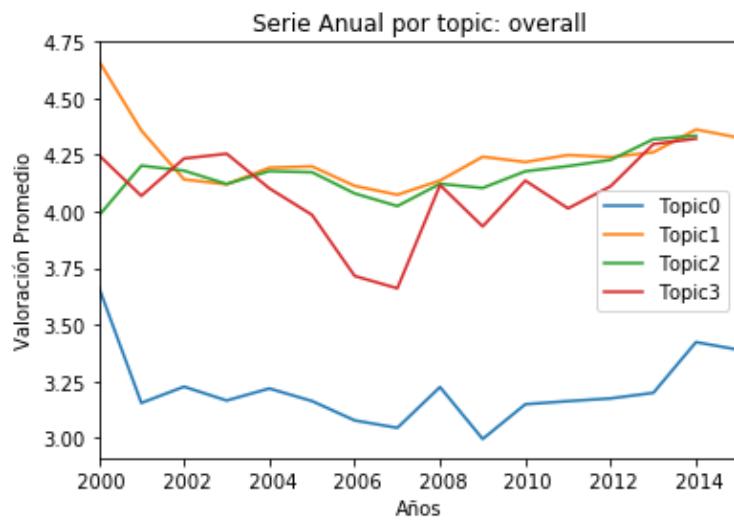


Figura 4.15: Series temporales: Valoración promedio de opiniones por año.

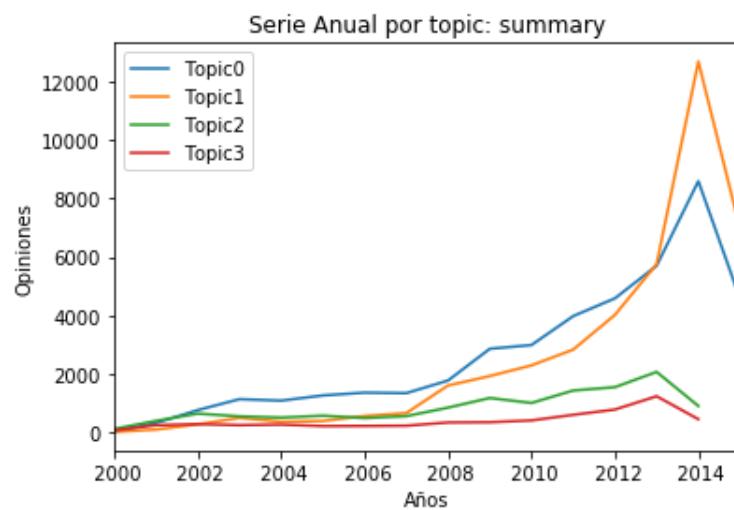


Figura 4.16: Series temporales: Cantidad de opiniones por año.

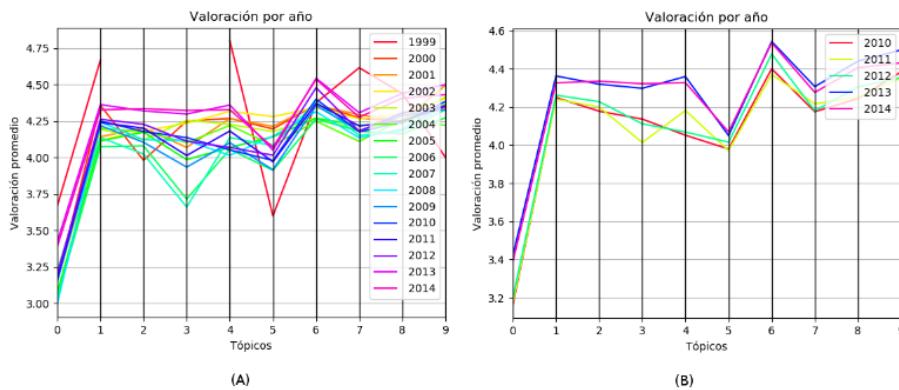


Figura 4.17: Parallel plot: Valoración de opiniones por año, a 16 años y b 5 años.

```

1 graficosTiempo= graphicsTime(dfVideoGames)
2 graficosTiempo.parallelPlot(2007,2014,target = 'Anos', xlabel = 'Topics', ylabel = 'Valoracion promedio', title = 'Valoracion por año')

```

Mapa competitivo

Buscando contrastar las opiniones de productos en el marco de variables categóricas se presenta una visualización que refleja la polaridad promedio de las opiniones acotado en un universo de 10 categorías de tópicos, como en la Figura 4.18 en a y b.

La librería de visualización del mapa competitivo se desarrolla mediante el método `barChartVs2`, este incluye un pre procesamiento gestionado mediante la función `ProductsToArray` que genera arreglos de datos con opiniones con polaridad positiva y negativa, que permite la representación en un plano cartesiano con centro en el eje vertical. Dichas polaridades no representan el sentimiento asociado, para este caso el único dato de interés es la magnitud. La Figura 4.17 a y b muestran un dominio de las opiniones del producto uno, que supera al producto 2 cubriendo tres tópicos más.

La librería de visualización del mapa competitivo se desarrolla mediante el método `barChartVs3`, este incluyen un pre procesamiento que genera arreglos de datos con opiniones con polaridad positiva e inversa mediante los métodos `competitiveSummary` y `ProductsToArray` para poder representar los datos en un plano cartesiano con centro en el eje vertical. Las polaridades no representan el sentimiento asociado, para este caso el único dato de interés es la magnitud. La Figura 4.17 a y b muestran un dominio de las opiniones del producto uno, que supera al producto 2 cubriendo tres tópicos más. Las Figuras 4.18 y 4.17 se generan con las siguientes líneas de código:

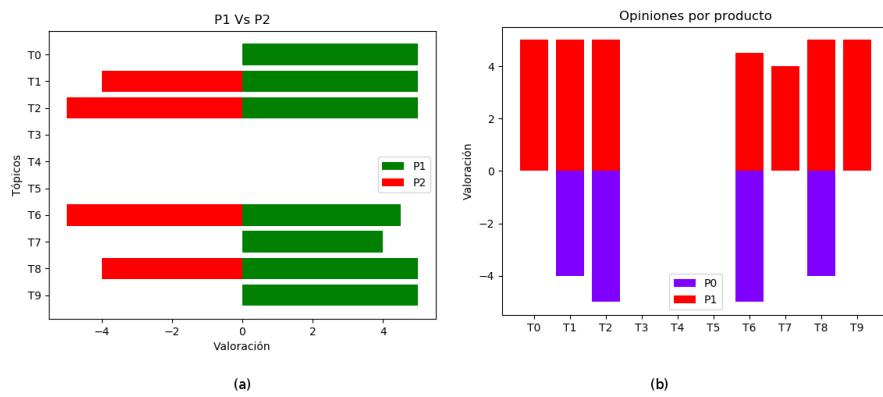


Figura 4.18: Mapa competitivo: P1 vs P2 por cada tópico, *a* horizontal, *b* vertical.

```

1 products = ['B000006POK', '6050036071']
2 graficosComparacion = graphicsCompare()
3 graficosComparacion.competitiveSummary(dfVideoGames, products)
4 graficosComparacion.productsToArray(products)
5 graficosComparacion.horizontalBar('Producto 1 Vs Producto 2')
6 graficosComparacion.barChartVs3(ylabel='Valoracion', Title= 'Opiniones
    por producto')

```

Observador de opiniones

La comparación de múltiples productos en distintos tipos de categorías se implementó combinando la cantidad de opiniones con su valoración promedio, para mostrar una medida ponderada de estas dos métricas. Cada producto es evaluado en el contexto de cada variable categórica como se muestra en la Figura 4.19, donde la categoría del eje horizontal es asociada a un tópico, mostrando un valoración de la importancia de las opiniones dentro de cada tema.

Se evidencia a grandes rasgos opiniones negativas en gran magnitud para el producto 2 en el tópico cero. El caso contrario ocurre para este mismo producto en el tópico nueve, donde consigue la puntuación máxima. La Figura 4.19 se genera con las siguientes líneas de código:

```

1 products = ['B000006POK', '6050036071', '0700099867']
2 graficosComparacion = graphicsCompare()
3 graficosComparacion.polaritySummary(dfVideoGames)
4 graficosComparacion.productsToArray(products)
5 graficosComparacion.barChartVs2(ylabel='Valoracion', Title= 'Opiniones
    por producto')

```

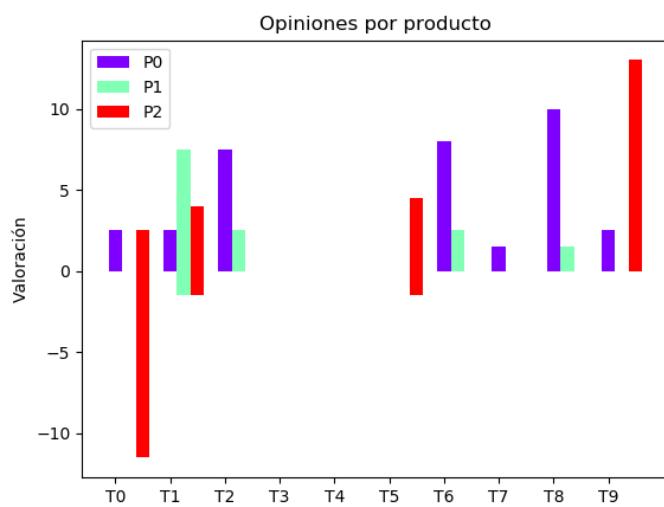


Figura 4.19: Observador de opiniones, análisis de tópicos de tres productos.

Capítulo 5

Caso de aplicación: Consolas de vídeo juegos

En este capítulo se implementará la librería desarrollada en un caso de aplicación real. Se realizará una descripción del conjunto de datos en la Sección 5.1, en la 5.2 se exponen los productos del análisis. A continuación en la sección 5.3 y 5.4 se desarrolla el análisis visual sobre el conjunto de datos y los productos seleccionados, entorno a los tópicos y características de las consolas de vídeo juego de salón y portables indicadas a continuación:

Se realizará un análisis de características y tópicos entorno a las consolas de vídeo juegos. Aquí se mostrarán algunas de las bondades y limitaciones de las técnicas de análisis visual. El análisis tendrá dos perspectivas, la primera dirigida al análisis de características y la segunda de tópicos. Los productos seleccionados son las consolas de vídeo juegos de bolsillo y salón de las marcas Sony; PSP, PSP Vita, PS2 y PS3, de Microsoft; Xbox y Xbox360, y Nintendo; el DSI y Wii. A continuación una breve descripción de los productos de análisis:

- Sony PSP: Por sus siglas PlayStation Portable es una vídeo consola de bolsillo lanzada en el año 2004. Su última generación se fabricó hasta el año 2014. Se registraron ventas de 80.82 millones en el 2015.
- PS Vita: Esta vídeo consola es la sucesora de la PSP, se lanzó en el año 2011 y continua vigente su fabricación y distribución. Esta generación registro 16 millones de dispositivos vendidos hasta marzo del 2018.
- Sony PS2: Consola de vídeo juegos de sobremesa de Sony lanzada en el año 2000, su fabricación fue descontinuada en 2013 registrando a la fecha 158 millones de consolas vendidas.
- Sony PS3: Es la sucesora del PS2, fue lanzada en el año 2006 y su producción finalizó en el año 2017, con un registro en ventas de 86 millones de consolas.

- Nintendo DSI: Es una video consola portátil lanzada en el año 2008 con 159 millones de unidades vendidas.
- Nintendo Wii: Consola de video juegos de salón lanzada en el año 2006 cuya producción se descontinuo en el año 2016. Se registraron un total de 101 millones de dispositivos vendidos.
- Microsoft Xbox: Consola de salón lanzada en el año 2002, se descontinuo su producción en el año 2008, con ventas 24 millones de unidades.
- Microsoft Xbox360: Sucesora del Xbox lanzada en el año 2005 que se descontinuo en el año 2016 registrando 85 millones de dispositivos.

Para el análisis el conjunto de datos de estudio cuenta con un histórico de 16 años comprendido desde mayo de 1996 hasta julio de 2014. Observando el tiempo de vida de los productos aparecen interesantes intersecciones a analizar, como es el caso de las consolas de sobre mesa Ps2 y Xbox a partir del año 2002. Xbox360, Wii y PS3 desde el 2006. Por último DSI, PSP desde el 2008 y la Vita desde el 2011. Teniendo el contexto anterior se plantea una comparativa de estos productos bajo distintas perspectivas para comprender la posición de sus clientes respecto a las características y tópicos entorno a los distintos comentarios dejados por los usuarios.

Inicialmente se observará el comportamiento de las distribuciones, visualizando las valoraciones con el diagrama de barras de la Figura 5.1, la representación de serie temporal de la Figura 5.4 y la visualización de tópicos con el stack plot de la Figura 5.2. Dichas distribuciones fueron planteadas filtrando los meta datos de cada producto según la clasificación que proporciona Amazon en el campo “categories”. En la práctica dicha clasificación mostró tener cierta incertidumbre, en especial en los campos de tiempo, donde se identifican inconsistencias con el tiempo de vida de algunos productos, donde aparecieron opiniones de productos con varios años de antelación a su lanzamiento. Un segundo aspecto a destacar es la poca relevancia de algunos comentarios que no se refieren al producto de manera directa. Por lo antes mencionado, los gráficos con excepción de los arriba reverenciados cuentan con un filtro adicional que tienen en cuenta solo los comentarios donde se hace mención explicita del producto.

5.1. Fuente de datos

Las técnicas de análisis visual adquieren relevancia desde varias dimensiones y contextos como vio en la Sección 2, dependiendo básicamente de dos parámetros: el objetivo de visualización y los datos de los que se dispone. Para efectos de este trabajo se dispondrá de un conjunto de datos de propiedad de la compañía de comercio electrónico Amazon, proporcionado por [27], que está disponible en <http://jmcauley.ucsd.edu/data/amazon/links.html>.

El conjunto de datos esta conformado por 142.8 millones de opiniones de productos, tomados desde mayo de 1996 hasta julio de 2014. Además de incluir el texto con cada opinión, se dispone también de votos sobre la utilidad de dicha opinión y meta datos relevantes con descripciones del producto, fechas, categorías, precios, marcas, imágenes y características.

Los productos están organizados en 24 categorías: “Books”, “Electronics”, “Movies and TV”, “CDs and Vinyl”, “Clothing”, “Shoes and Jewelry”, “home and kitchen”, “kindle store”, “sports and outdoors”, “cell phones and accessories”, “health and personal care”, “toys and games”, “tools and home improvement”, “beauty”, “apps for android”, “office products”, “pet supplies”, “automotive”, “grocery and gourmet food”, “patio, lawns and garden”, “baby”, “digital music”, “musical instruments”, y “amazon instant video”. La fuente provee también grupos de sub conjuntos para experimentación, donde los datos han sido filtrados con distintos criterios.

El conjunto de datos completo “raw review data” tiene un tamaño de 20 Gb y contiene 142.8 millones de opiniones, “user review data” viene también en un formato de 18Gb donde se ha removido ítems duplicados, dejando 83.68 millones de opiniones. Este mismo conjunto de datos se encuentra en “product review data”, donde se ha ordenado por producto. Una versión más simple está en “ratings only” de 3.2 Gb que contiene solo la evaluación del producto. “5-core” de 9.9 Gb contiene un sub conjunto con todas las opiniones de productos que tienen mínimo 5 calificaciones. Finalmente se dispone de “aggressively deduplicated data” de 18 Gb donde las opiniones repetidas se han eliminado sin importar si provienen de distintos clientes.

El formato de datos de ejemplo se ilustra en el JSON¹ a continuación:

```
1 { "reviewerID": "A2SUAM1J3GNN3B", "asin": "0000013714", "reviewerName": "J. McDonald", "helpful": [2, 3], "reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!", "overall": 5.0, "summary": "Heavenly Highway Hymns", "unixReviewTime": 1252800000, "reviewTime": "09 13, 2009" }
```

donde:

- reviewerID - ID de el cliente que opinó, e.g. A2SUAM1J3GNN3B
- asin - ID de el producto, e.g. 0000013714
- reviewerName - nombre del cliente
- helpful - ratio de utilidad de la opinión, e.g. 2/3
- reviewText - texto de la opinión

¹ Acrónimo de JavaScript Object Notation, es un formato de texto ligero para el intercambio de datos.

- overall - ratio del producto
- summary - resumen de la opinión
- unixReviewTime - tiempo del comentario (unix time)
- reviewTime - tiempo del comentario (raw)

A continuación un conjunto de ejemplo de meta datos:

```
1 { "asin": "0000031852", "title": "Girls Ballet Tutu Zebra Hot
Pink", "price": 3.17, "imUrl": "ecx.imagesamazon.com/images/
I/51fAmVkBbyL.SY300.jpg", "related": \{ "also_bought": [
"BO0JHONN1S", "B002BZX8Z6", "B00D2K1M30", "0000031909", "B00613WDTQ
"], "bought_together": ["B002BZX8Z6"] \}, "salesRank": \{
"Toys & Games": 211836 \}, "brand": "Coxlures", "categories": [
["Sports & Outdoors", "Other Sports", "Dance"] ] }
```

donde:

- asin - ID del producto, e.g. 0000031852
- title - nombre del producto
- price - precio en dolares (precio de cuando fue comprado)
- imUrl - url de la imagen del producto
- related - productos relacionados (vistos y/o comprados juntos)
- salesRank - Ratio de ventas
- brand - Nombre de la marca
- categories - categorías

Ciertos tipos de productos por el contexto tecnológico contienen una distribución uniforme en el tiempo, como el lector de libros digitales de Amazon “kindle” que fue lanzado en 2007. Es por esto que el análisis de este trabajo se centra sobre solo la categoría “video games” que contiene aspectos interesantes para destacar y contrastar las propiedades de ciertos tipos de gráficos.

5.2. Selección de productos

La base de datos no tiene delimitado claramente cada uno de los productos con su marca y referencia específica, los campos destinados para dicho fin contienen una alta cantidad de valores nulos. La selección de productos y características se realizó en una primera fase con apoyo del contenido en

los campos “categories” y “brand” de VideoGames5_metadata.json.gz y en una segunda de refinamiento con “reviewText” de reviews_Video_Games_5.

De los campos “categories” y “brand” se extrajeron el nombre de cada producto con su frecuencia asociada, donde los productos más relevantes en el contexto de vídeo consolas aparecieron:

- Xbox 360: 46657,
- PlayStation 3: 42143
- Wii: 23523, 'PlayStation 2': 17069
- Nintendo DS: 16882
- Nintendo 3DS: 10705
- Xbox: 7084,
- GameCube: 5468,
- PlayStation 4: 5448
- Wii U: 5408
- Sony PSP: 5299
- PlayStation Vita: 4990
- Xbox One: 4677
- PlayStation: 3975

La frecuencia de productos mostrada se generó con las siguientes líneas de comando:

```
1 categoriasArray = [list(chain(*categoria)) for categoria in dfMeta.
2     categorias]
3 categoriasLista = list(chain(*categoriasArray))
4 categoriasFrecuencia = Counter(categoriasLista)
5 categoriasFrecuencia.most_common(100)
```

Los productos a incluir en el análisis se filtran cada uno en un conjunto de datos nuevo. A continuación se aplica un nuevo análisis sobre el campo “reviewText” para identificar en cada opinión las diferentes formas de referirse a un producto en el espectro de opiniones. Una vez reconocidos estos se aplica un filtro para extraer solo las opiniones que se refieren en específico al producto de estudio. El paso anterior se realiza con cada producto individualmente.

5.3. Análisis de características y tópicos

Para la extracción de tópicos, se implementó el modelo LDA entrenado para 10 tópicos que contiene el conjunto de palabras que representan cada documento, cada una de estas palabras tiene una probabilidad asociada, como se muestra a continuación:

1. (0, 0.061*34 + 0.025*headset + 0.016*8217 + 0.014*dlc + 0.014*song + 0.012*danc + 0.008*onlin + 0.006*team + 0.005*improv + 0.005*featur)
2. (1, 0.013*map + 0.013*mission + 0.011*multiplay + 0.010*campaign + 0.006*singl + 0.006*build + 0.006*onlin + 0.006*battlefield + 0.005*call + 0.005*base)
3. (2, 0.023*xbox + 0.015*ps3 + 0.013*consol + 0.013*button + 0.011*360 + 0.009*vita + 0.008*charg + 0.007*screen + 0.007*batteri + 0.007*tv)
4. (3, 0.020*weapon + 0.015*gun + 0.014*zombi + 0.014*kill + 0.012*shooter + 0.011*shoot + 0.011*dead + 0.010*op + 0.009*cod + 0.007*action)
5. (4, 0.022*app + 0.016*sonic + 0.015*soul + 0.013*guitar + 0.011*item + 0.009*quest + 0.008*skill + 0.008*hero + 0.008*dark + 0.008*diablo)
6. (5, 0.040*wii + 0.032*3d + 0.028*mario + 0.023*u + 0.022*nintendo + 0.015*super + 0.009*zelda + 0.008*version + 0.008*collect + 0.008*kart)
7. (6, 0.005*jump + 0.005*batman + 0.005*button + 0.004*attack + 0.004*area + 0.004*hit + 0.004*open + 0.003*object + 0.003*boss + 0.003*place)
8. (7, 0.007*battl + 0.007*experi + 0.006*rpg + 0.006*voic + 0.006*combat + 0.005*interest + 0.005*final + 0.005*quest + 0.004*titl + 0.004*develop)
9. (8, 0.010*bought + 0.009*price + 0.009*kid + 0.007*money + 0.007*son + 0.007*wait + 0.006*purchas + 0.006*figur + 0.005*thought + 0.005*amazon)
10. (9, 0.021*ps4 + 0.013*mous + 0.009*pc + 0.009*car + 0.008*issu + 0.008*drive + 0.008*updat + 0.008*race + 0.007*xbox + 0.007*download)

La lista de palabras por tópico se generó mediante las siguientes instrucciones de código:

```
1 ldamodel.show_topics(num_topics=10, num_words=10)
```

El proceso de etiquetado de cada ítem es riguroso y requiere de un análisis exhaustivo, donde la interpretación puede llegar a tener un grado de subjetividad elevado. Al estudiar las probabilidades más altas en cada tópico y la relación de las palabras dentro de un determinado contexto se determinaron las siguientes etiquetas para cada tema:

1. Sonido: headset, song,dance, music con probabilidades entre el 2,5 y 0.5 porciento.
2. Características de vídeo juegos: Map, missions, multiplayer, campain, online, single con probabilidades entre el 1.3 y 0.5 porciento.
3. Características consolas vídeo juegos: xbox, ps4, console, 360, vita, button, charge, screen con probabilidades entre el 2.3 y 0.7 porciento.
4. Vídeo juegos de disparo: Weapon, gun, zombi, kill, shooter, shoot,dead con probabilidades entre el 2.1 y 1.3 porciento.
5. Desconocido.
6. Juegos clásicos de Nintendo: Wii, mario, selda, nintendo, kart con probabilidades entre el 4 y 0.08 porciento.
7. Desconocido.
8. Juegos de guerra y RPG ²: battle, rpg, combat con probabilidades entre el 7 y 6 porciento.
9. Aspectos financieros: bought, price, money, purchase, amazon con probabilidades entre el 10 y 5 porciento.
10. Juegos de conducción: Car, drive, race con probabilidades entre el 9 y 8 porciento.

Los temas identificados como “Desconocido” representaron cierta dificultad en el etiquetado ya que a simple vista no mostraron estar alineadas con algún tema conocido o de fácil identificación.

Cada producto tiene un conjunto de datos asignado. Cada uno será sometido a análisis para identificar los sustantivos que hacen referencia a los atributos del producto. Al final el objetivo de este paso es definir un diccionario de sinónimos que permita reconocer cada característica, como el que se muestra a continuación:

- conford: buttons, button, fingers, handheld, feel y feels.

²Es un género de videojuegos que usa elementos de los juegos de rol tradicionales.

- gráficos: graphics, pixels, screen, screens y resolution.
- memoria: sd, card, memory y gb.
- batería: battery y hours.
- juegos: game, games y titles.
- precio: price, money y cost.
- software: system, systems, settings y controller.

La extracción de tópicos se desarrolla sobre el campo “maximos” que contiene las etiquetas de los tópicos de cada opinión. La extracción se hizo mediante el algoritmo LDA como se describió en las Secciones 2.2.1 y 4.1.2. Adicional a este proceso se implementó una rutina que crea un mapa para relacionar los códigos de producto “asin” con su respectivo producto y tópico.

La distribución de las valoraciones permite identificar la estadística del conjunto de productos y sus categorías. Dentro de este análisis se pretende visualizar la valoración media del producto enfocado en los 10 tópicos definidos anteriormente.

5.3.1. Consolas de salón

Visualizando la cantidad de opiniones por cada categoría de valoración, se observa una pareja competencia entre el Xbox 360 y la PS3, donde el dispositivo de Microsoft saca una ligera ventaja en la categoría “Exelente”. En términos generales se puede observar una clara mayoría en las valoraciones con etiqueta Bueno y Exelente.

El Stack plot muestra interesantes propiedades para visualizar cantidades de variables categóricas, la Figura 5.2 compara el histórico de opiniones de dos consolas, Xbox y Play Station 2, que han sido contrastados contra un dispositivo de una generación más reciente, el Nintendo Wii. Este a pesar de que se empezó a vender 5 años después del lanzamiento de Ps2 y el Xbox, casi que duplico en número de opiniones.

Realizando una comparación del Nintendo Wii con sus rivales contemporáneos el Xbox 360 y la Play Station 3 se aprecia una cantidad de opiniones muy inferior a sus competidores. Parece existir una comunidad de clientes de Amazon más activa entorno a el Xbox 360 seguido por la Play Station 3, como se puede observar en la Figura 5.3

La serie temporal de la Figura 5.4 muestra un claro aumento de la cantidad de opiniones y valoración promedio de las mismas en la medida que el tiempo aumenta. En promedio los niveles de valoración promedio son similares, la única diferencia relevante es el número inferior de opiniones de la consola Nintendo Wii. A Partir del año 2013 la cantidad de opiniones

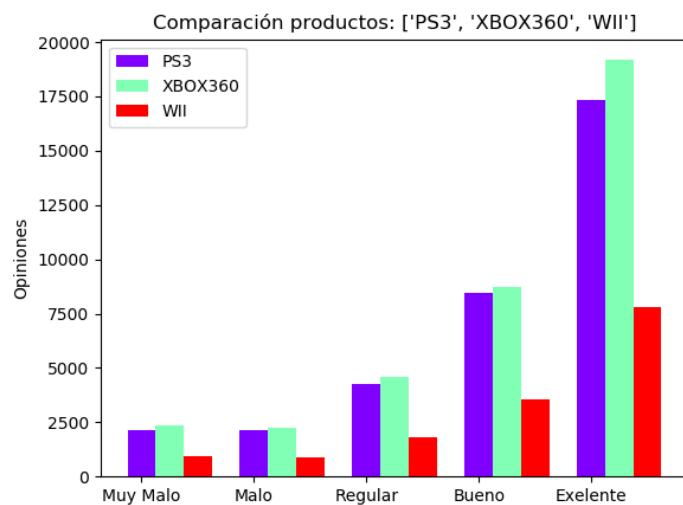


Figura 5.1: Diagrama de barras, comparación de categorías de valoración de clientes.

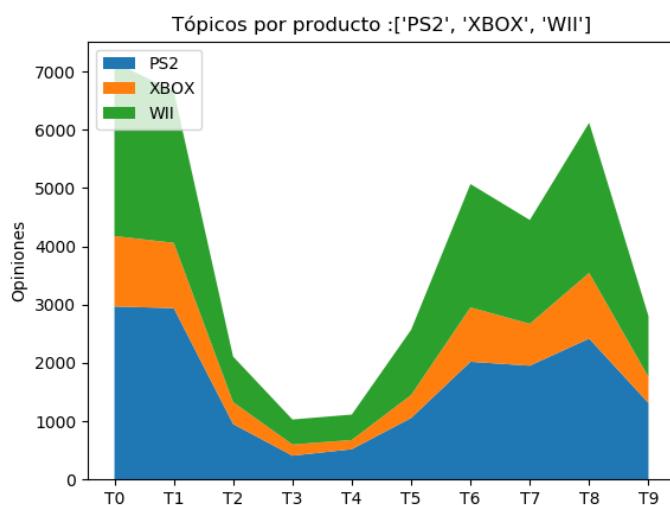


Figura 5.2: Stack plot, comparación productos Nintendo Wii con las generaciones anteriores de sus competidores, las consolas Xbox y Play Station 2.

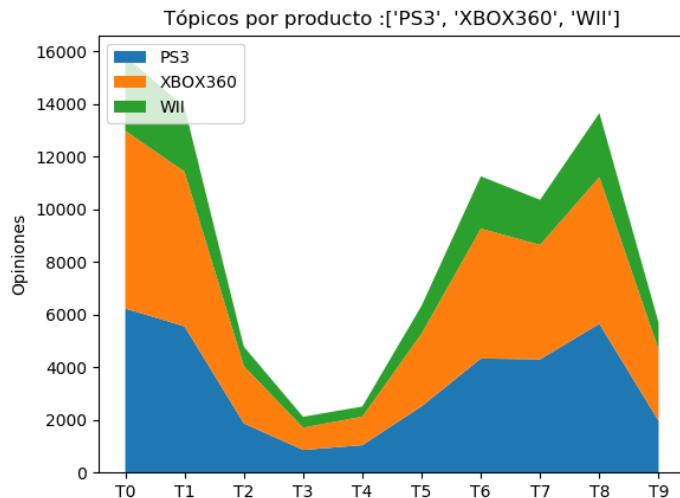


Figura 5.3: Stack plot, comparación productos Nintendo Wii y sus competidores, las consolas Xbox 360 y Play Station 3.

decrece bruscamente, fecha que coincide con el lanzamiento de las nuevas generaciones de estos dispositivos: Nintendo Wii U a finales del 2012 y Play Station 4 con la Xbox One en 2013.

Una visualización tipo Pie permite obtener una visión global de la distribución de variables categóricas. En las Figuras 5.5,5.6 y 5.7 se observan dos dimensiones, la primera en color representa la distribución de tópicos, la segunda muestra la valoración promedio. El tópico T1 contiene la mayoría de las opiniones que se refieren a configuraciones multijugador y de conectividad, donde la consola Nintendo Wii en la Figura 5.5 muestra una valoración más baja respecto a sus competidores.

En cuanto al aspectos de sonido representado por T0, las tres consolas muestran valoraciones bajas, en especial la consola de Nintendo, que registra la peor calificación que se puede obtener. Algunos temas parecen existir uniformidad, como es el caso de T8 que se refiere a aspectos monetarios como el precio, o el T7 que representa temas entorno a los juegos de guerra y RPG.

El tópico T3 que se refiere a juegos de disparo parece mostrar como factor común una baja valoración, en especial en la consola Xbox360 como se aprecia en la Figura 5.7.

La consola Nintendo Wii entró al mercado en el 2006 convirtiéndose así competidora directa de la Play Station 3 y el Xbox 360. En la Figura 5.8 se muestra una métrica que multiplica el número de opiniones por la valoración promedio otorgada por el cliente de dos productos, el Play Station

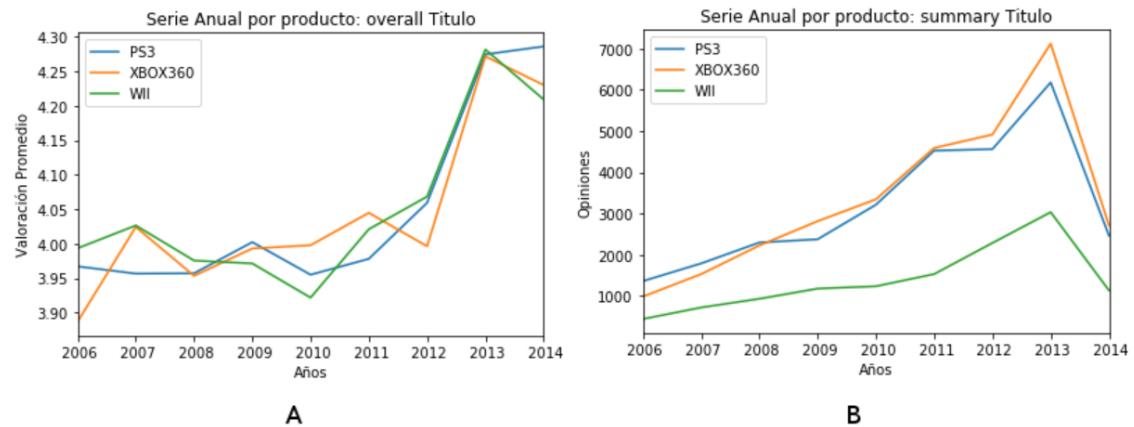


Figura 5.4: Serie temporal productos Xbox 360, PS3 y Wii. a)Valoraciones promedio, b) Cantidad de opiniones.



Figura 5.5: Diagrama de Pie Consola Nintendo Wii.



Figura 5.6: Diagrama de Pie consola de Sony: PS3.



Figura 5.7: Diagrama de Pie consola de Microsoft: Xbox 360.

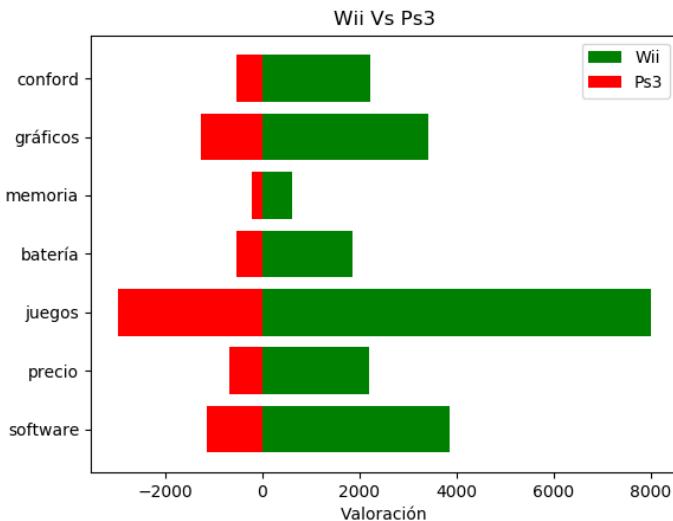


Figura 5.8: Stack plot, comparación productos Nintendo Wii con las generaciones anteriores de sus competidores, las consolas Xbox y Play Station 2.

3 y Nintendo Wii. Las dos consolas estuvieron en el mercado cerca de una década. El Nintendo Wii destaca claramente en la métrica propuesta, esta revela una mayor cantidad de opiniones positivas.

Los tópicos de la Figura 5.9 A y B representan 3 de los principales tópicos encontrados, tópico 2 (en azul) características de conectividad y multijugador, tópico 3 (en naranja) características del dispositivo y tópico 9 (en verde) precios. La representación temporal muestra para Xbox unos picos de calificaciones bajas durante los años 2007 y 2008 para los tópicos 2 y 9. Un año después, durante el 2008 y 2009 la calificación aumenta considerablemente. Los años restantes la serie muestra valores con una media superior al 4.

La Play Station 2 parece mejorar considerablemente la experiencia de multijugador y sus precios desde el año 2012, curiosamente durante su época de jubilación en el mercado.

La implementación del Opinion observer sobre el PS3, Xbox 360 y Wii se observa en la Figura 5.10. Allí T0, el primer tópico hace referencia al audio en la consola, y el segundo T1, a características multijugador y online. Estas muestran una métrica favorable para el Nintendo Wii en la medida que supone una menor cantidad de opiniones negativas. En la misma T1 se aprecia un destacable número de opiniones positivas del Xbox 360 y PS3. El tópico T8 que se refiere a los precios de las consolas favorece ligeramente al PS3 que muestra una considerable cantidad de opiniones positivas.

Los juegos de conducción representados con tópico T9 muestran una clara ventaja en la consola de Nintendo que duplica en opiniones positivas

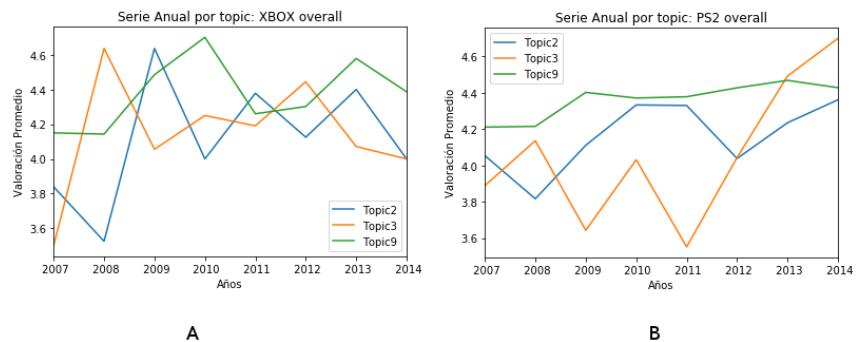


Figura 5.9: Serie temporal, comparación de tópicos productos a)Xbox y b) Play Station 2.

a sus competidores.

El Parallel plot muestra dinámicamente los cambios y saltos de la variable categórica tópicos en la dimensión tiempo y valoración promedio. Los juegos de disparo representados por el tópico 3 parecen mostrar una ligera disminución en los años 2010, 2012, 2006 y 2007. Algo similar sucede con el tópico T0, donde la valoración promedio en todos los años se muestra por debajo de 4. Los tópicos del 5 al 9 muestran calificaciones con menos variabilidad, todas encima de una puntuación de 4.

5.4. Análisis de características

El análisis de características se realizó sobre los atributos vistos al inicio de esta sección, donde el objetivo aquí es contrastar consolas de video juegos de salón y portable de una misma generación. La extracción de características se realizó como se describe en la Sección 4.1.3, implementando filtros iterativos sobre corpus de textos para extraer los sustantivos de interés.

5.4.1. Consolas de salón

En la gama de consolas de salón, se postulan para el análisis nuevamente las consolas Xbox 360, PS3 y Nintendo Wii. El Xbox 360 mostró un menor promedio de opiniones negativas en los atributos memoria y batería. En rasgos generales el comportamiento de las valoraciones es en promedio similar para las consolas Wii y PS3, como se ilustra en la Figura 5.12.

Como muestra la Figura 5.13, al analizar las valoraciones ponderadas por la cantidad de opiniones, el PS3 destaca con una valoración ponderada positiva en todo el rango. El Xbox 360 destaca en el atributo juegos, donde muestra una cantidad inferior de valoraciones negativas. En cuanto a memoria la PS3 se lleva la mejor valoración superando ampliamente a

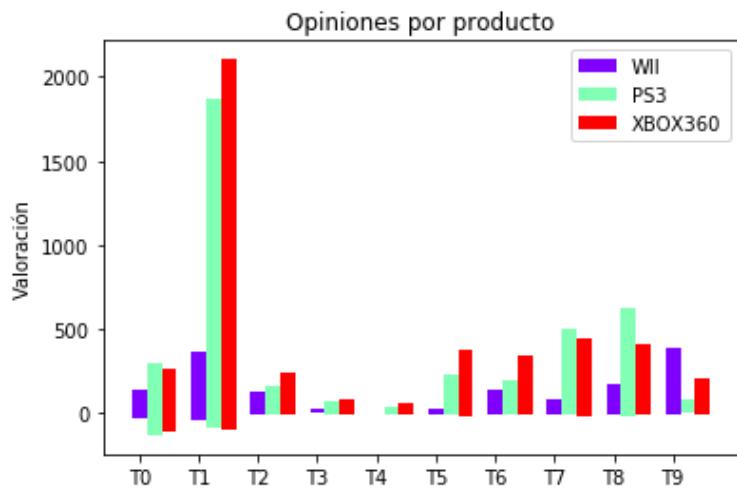


Figura 5.10: Representación de tópicos con el Opinion observer para Xbox 360, PS3 y Wii.

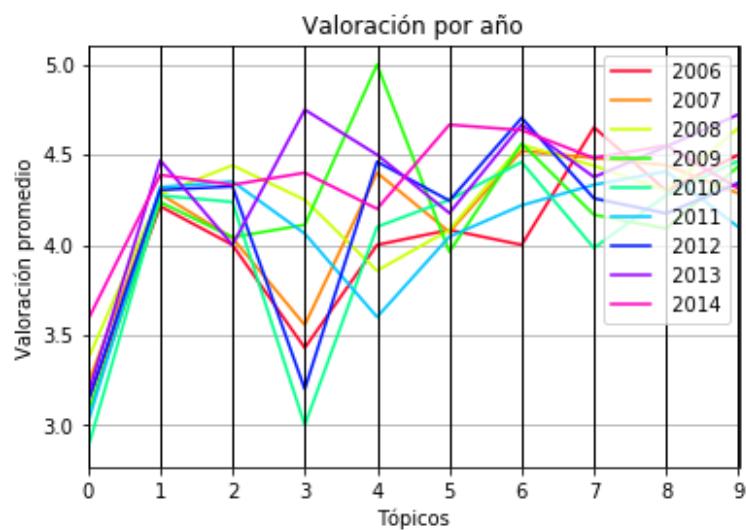


Figura 5.11: Representación del histórico de tópicos con el Parallel plot sobre el total de opiniones de Xbox 360, PS3 y Wii.

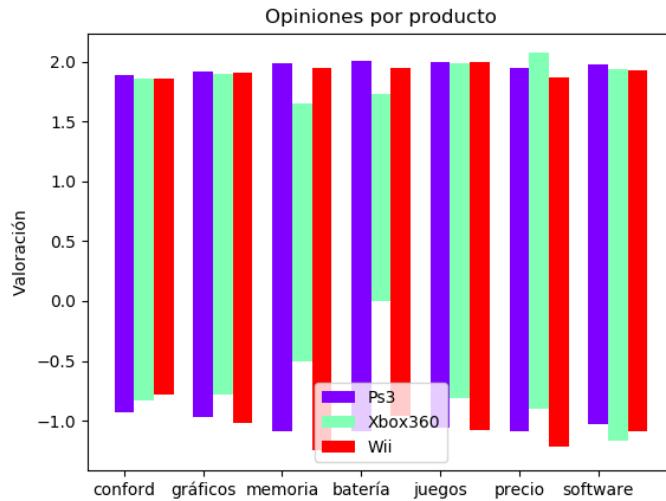


Figura 5.12: Representación de valoración promedio de características con el Opinion observer para Xbox 360, PS3 y Wii.

sus competidores. De forma general el PS3 y el Wii parecen mostrar una valoración ponderada muy similar.

5.4.2. Consolas portables

Las consolas portátiles representan distintas generaciones de sus líneas originales de producto, así la Play Station Vita es la continuación de la PSP que es competidora directa de la consola portable Nintendo DSI. La valoración promedio de la Figura 5.14 muestra puntos débiles respecto al conford de la consola Ps Vita. En cuanto al atributo juegos, la consola de Nintendo saca ventaja a sus competidores, exhibiendo una cantidad promedio de opiniones, ligeramente superior y una clara ventaja debido a la mínima cantidad de opiniones negativas que registra. Un caso muy similar sucede con el precio, donde la DSI vuelve a destacar. En temas de memoria es donde parecen compartir más elementos en común, mostrando una valoración promedio muy similar todos los productos.

Al tener en cuenta la cantidad de opiniones en la representación de la Figura 5.15, la consola Play Station Vita destaca notablemente sobre el resto. La PSP muestra un comportamiento muy parejo en todos sus atributos, con una valoración promedio cercana a 30.

Al comparar mediante el Competitive Char los productos DSI y PSP se aprecian ligeras diferencias mostrando el DSI una mejor valoración de su memoria y una leve ventaja sobre la batería del PSP, como se ilustra en la

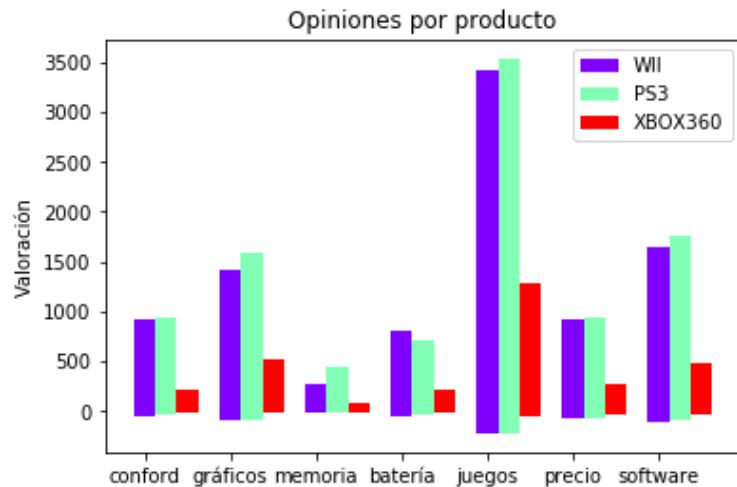


Figura 5.13: Representación de valoración ponderada de características con el Opinion observer para Xbox 360, PS3 y Wii.

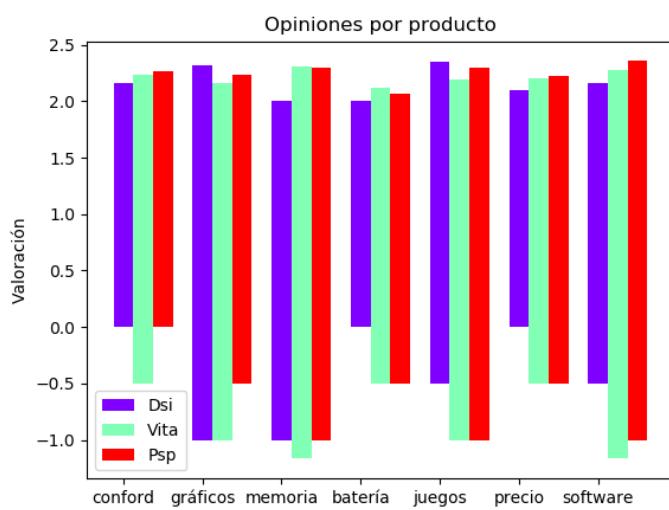


Figura 5.14: Representación de valoración promedio de características con el Opinion observer para DSI, Vita y PSP.

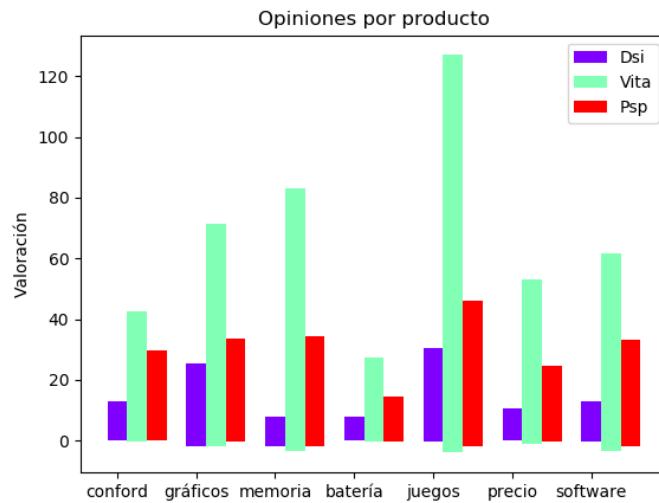


Figura 5.15: Representación de valoración ponderada de características con el Opinion observer para DSi, Vita y PSP.

Figura 5.16

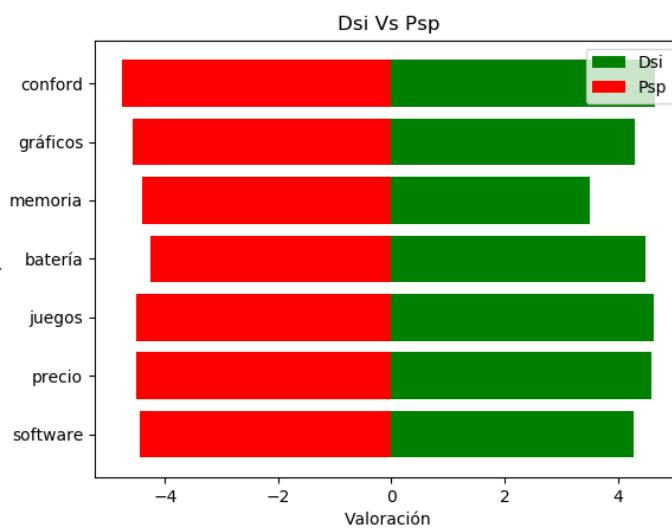


Figura 5.16: Representación de valoración promedio de características con el Competitive Chart para DSI y PSP.

Capítulo 6

Conclusiones y trabajo futuro

Las técnicas de visualización estudiadas mostraron distintas propiedades que favorecen el análisis en determinados contextos. La comprensión y el estudio a profundidad de dichas propiedades proporcionan la base para que se aproveche el potencial de cada técnica. La implementación de la librería desarrollada expone interesantes características, que llevan los datos de opiniones de usuario a un nivel de abstracción superior. Por otro lado el análisis realizado sobre las técnicas de visualización, deja en evidencia claras oportunidades de mejora de algunas de las técnicas existentes, y plantea la posibilidad de explorar nuevas áreas en el campo de visualización de opiniones.

Se presento en la Sección 3, un análisis de las técnicas de visualización entorno a la minería de opiniones. La Sección 4 expuso la construcción de una librería de visualización de opiniones en Python. Por último, la Sección 5 desarrolló un caso de aplicación real de la librería construida, sobre un conjunto de datos de propiedad de la compañía de comercio electrónico Amazon. El resultado final del caso de aplicación, se materializa en un análisis de las características y tópicos entorno un grupo de consolas de video juegos. De este trabajo hemos extraído las siguientes conclusiones:

- Se identificaron dos principales restricciones de las técnicas de visualización para el análisis de opiniones: En primer lugar la cantidad de datos que pueden representar, esto debido a todos los procesos de procesamiento, en especial los que incluyen procesamiento del lenguaje natural, dichas rutinas tienen un costo computacional alto. El segundo elemento es la cantidad de dimensiones que se pueden representar, a pesar de los intentos de mostrar más de 4 dimensiones en un mismo gráfico, las figuras se tienden a sobre saturar, haciendo que disminuya la claridad del mensaje que se transmite, como es el caso del Opinion Ring.

- Los métodos de extracción de tópicos y características no son cien por ciento automáticos, se apoyan en procesos de etiquetado humano, para interpretar el significado de cada tópico y definir el contexto de palabras (sustantivos con significado intrínseco o extrínseco) que definen un atributo en la extracción de características. Todo esto hace que se incorpore un alto nivel de subjetividad en la construcción de sistemas para el análisis de opiniones.
- La normalización de los ejes de los gráficos permite tener una concepción más adecuada de las proporciones de los objetos de estudio, pero es evidente que hace que se pierda la noción de magnitud. Por lo que es conveniente en muchos casos (como el Opinion Observer) visualizar los gráficos con y sin normalización.
- El nivel de granularidad de los datos puede ser complejo de conseguir, pero es importante para evitar representaciones con sesgos debido al des-balanceo de las clases de los datos que representa. Así por ejemplo una característica puede estar compuesta por dos sub características más, y si la cantidad de datos de una de las dos clases es muy superior a la otra puede originar representaciones sesgadas hacia la clase mayoritaria. Teniendo en cuenta lo anterior, lo más fiel sería representar individualmente las sub características o implementar un sistema de pesos que represente la característica. Lo planteado anteriormente aplica también en la granularidad que se puede obtener para las marcas de los productos con sus distintos modelos, diferentes generaciones y versiones de estos.
- Visualizaciones de alta dimensionalidad y gran volumen de datos pueden llegar a adaptarse mejor a las técnicas existentes, incluyendo etapas de pre procesamiento que resuman la información. Esto a su vez tiene un coste, ya que se pierde granularidad en la información y en dados casos puede que se requiera del apoyo de un experto.

Como trabajo futuro, se propone implementar procesos de análisis de sentimientos, en vista que muchas fuentes de datos no disponen de valoración de los clientes que permita inferir este aspecto. La implementación de librerías más eficientes es otro tema de trabajo futuro, donde se plantea migrar el desarrollo basado en la librería de procesamiento del lenguaje natural “NLTK” de Python a un paquete más eficiente como “spaCy” o a un entorno mejor preparado para gestionar grandes volúmenes de datos como Pyspark. Igualmente en cuanto a representación, existe una gran cantidad de trabajo para construir las técnicas de visualización de opiniones, sobre otras estructuras como diagramas de burbujas ó diagramas de Venn.

Bibliografía

- [1] André Luiz Firmino Alves, Cláudio de Souza Baptista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva. A spatial and temporal sentiment analysis approach applied to twitter microtexts. *Journal of Information and Data Management*, 6(2):118, 2016.
- [2] Kamal Amarouche, Houda Benbrahim, and Ismail Kassou. Product opinion mining for competitive intelligence. *Procedia Computer Science*, 73:358–365, 2015.
- [3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
- [4] Eivind Bjørkelund, Thomas H Burnett, and Kjetil Nørvåg. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, pages 229–238. ACM, 2012.
- [5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [6] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.
- [7] SRK Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, 34:569–603, 2009.
- [8] Anne Bruseberg and Deana McDonagh-Philp. Focus groups to support the industrial/product designer: a review based on current literature and designers' feedback. *Applied ergonomics*, 33(1):27–38, 2002.
- [9] Fernando Calderon, Chun-Hao Chang, Carlos Argueta, Elvis Saravia, and Yi-Shin Chen. Analyzing event opinion transition through summarized emotion visualization. In *Proceedings of the 2015 IEEE/ACM*

- International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 749–752. ACM, 2015.
- [10] Michael Chau and Jennifer Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1):57–70, 2007.
 - [11] Chaomei Chen, Fidelia Ibekwe-SanJuan, Eric SanJuan, and Chris Weaver. Visual analysis of conflicting opinions. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 59–66. IEEE, 2006.
 - [12] Yang Chen. Visual opinion analysis of threaded discussions. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 646–651. IEEE, 2015.
 - [13] Delenn Chin, Anna Zappone, and Jessica Zhao. Analyzing twitter sentiment of the 2016 presidential candidates. *News & Publications: Stanford University*, 2016.
 - [14] Dr Ciravegna et al. Adaptive information extraction from text by rule induction and generalisation. 2001.
 - [15] Weiwei Cui, Shixia Liu, Zhuofeng Wu, and Hao Wei. How hierarchical topics evolve in large text corpora. *IEEE transactions on visualization and computer graphics*, 20(12):2281–2290, 2014.
 - [16] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.
 - [17] Luigi Di Caro and Matteo Grella. Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5):442–453, 2013.
 - [18] Wenwen Dou, Isaac Cho, Omar ElTayeby, Jaegul Choo, Xiaoyu Wang, and William Ribarsky. Demographicvis: Analyzing demographic information based on user generated content. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pages 57–64. IEEE, 2015.
 - [19] Xiaolin Du, Yunming Ye, Raymond YK Lau, Yueping Li, and Xiaohui Huang. Multi-opinion ring: visualizing and predicting multiple opinion orientations in online social media. *Multimedia Tools and Applications*, 75(12):7159–7186, 2016.
 - [20] Fatima Zohra Ennaji, Abdelaziz El Fazziki, Mohamed Sadgal, and Djamal Benslimane. Social intelligence framework: Extracting and

- analyzing opinions for social crm. In *Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of*, pages 1–7. IEEE, 2015.
- [21] Espinosa, Jonathan. <https://github.com/jespinosal/visuaReviewsAnalysis>, enero de 2018.
 - [22] Larissa A Freitas and Renata Vieira. Ontology based feature level opinion mining for portuguese reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 367–370. ACM, 2013.
 - [23] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pages 121–132. Springer, 2005.
 - [24] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis. *Entropy*, 17:252, 2009.
 - [25] Michelle L Gregory, Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 23–30. Association for Computational Linguistics, 2006.
 - [26] Ming C Hao, Daniel A Keim, Umeshwar Dayal, and Jörn Schneidewind. Business process impact visualization and anomaly detection. *Information Visualization*, 5(1):15–27, 2006.
 - [27] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.
 - [28] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
 - [29] Minqing Hu and Bing Liu. Opinion feature extraction using class sequential rules. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 61–66, 2006.
 - [30] Alison Huettner and Pero Subasic. Fuzzy typing for document management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pages 26–27, 2000.

- [31] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.
- [32] Daisuke Ikeda, Hiroya Takamura, Lev-Arie Ratinov, and Manabu Okumura. Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.
- [33] Halldór Janetzko, Dominik Jäckle, and Tobias Schreck. Geo-temporal visual analysis of customer feedback data based on self-organizing sentiment maps. *International Journal on Advances in Intelligent Systems*, 7(1/2):237–246, 2014.
- [34] Wei Jin, Hung Hay Ho, and Rohini K Srihari. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1195–1204. ACM, 2009.
- [35] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.
- [36] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.
- [37] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.
- [38] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting evaluative expressions for opinion extraction. In *International Conference on Natural Language Processing*, pages 596–605. Springer, 2004.
- [39] Kostiantyn Kucher, Teri Schamp-Bjerede, Andreas Kerren, Carita Paradis, and Magnus Sahlgren. Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization*, 15(2):93–116, 2016.
- [40] Akshi Kumar, Prakhar Dogra, and Vikrant Dabas. Emotion analysis of twitter using opinion mining. In *Contemporary Computing (IC3), 2015 Eighth International Conference on*, pages 285–290. IEEE, 2015.

- [41] Zhichao Li, Min Zhang, Shaoping Ma, Bo Zhou, and Yu Sun. Automatic extraction for product feature words from comments on the web. In *Asia Information Retrieval Symposium*, pages 112–123. Springer, 2009.
- [42] Kar Wai Lim and Wray Buntine. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1319–1328. ACM, 2014.
- [43] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.
- [44] Yanping Lv, Xiao Xiao, Dazhen Lin, and Donglin Cao. Public opinion analysis based on geographical location. In *Image and Signal Processing (CISP), 2015 8th International Congress on*, pages 1215–1219. IEEE, 2015.
- [45] Zhiqiang Ma, Wenwen Dou, Xiaoyu Wang, and Srinivas Akella. Tag-latent dirichlet allocation: Understanding hashtags and their relationships. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 260–267. IEEE Computer Society, 2013.
- [46] Carla Ruiz Mafé and Silvia Sanz Blas. Influencia de las motivaciones en la decisión de compra y en la lealtad hacia internet. *Investigaciones europeas de dirección y economía de la empresa*, 12(3):195–215, 2006.
- [47] Wes McKinney. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, pages 1–9, 2011.
- [48] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, pages 101–110. ACM, 2008.
- [49] Qingliang Miao, Qiudan Li, and Ruwei Dai. Amazing: A sentiment mining and retrieval system. *Expert Systems with Applications*, 36(3):7192–7198, 2009.
- [50] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

- [51] Samaneh Moghaddam and Martin Ester. On the design of lda models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 803–812. ACM, 2012.
- [52] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349. ACM, 2002.
- [53] Finn Årup Nielsen. Wikipedia research and tools: Review and comments. *Browser Download This Paper*, 2012.
- [54] Rajdeep Niyogi et al. Demographic analysis of twitter users. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pages 2662–2667. IEEE, 2016.
- [55] NLProcessor – Text Analysis Toolkit. <http://www.infogistics.com/textanalysis.html>, 2000.
- [56] Daniela Oelke, Ming Hao, Christian Rohrdantz, Daniel A Keim, Umeshwar Dayal, Lars-Erik Haug, and Halldór Janetzko. Visual opinion analysis of customer feedback data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 187–194. IEEE, 2009.
- [57] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.
- [58] Nishith Pathak, Colin DeLong, Arindam Banerjee, and Kendrick Erickson. Social topic models for community extraction. In *The 2nd SNA-KDD workshop*, volume 8, page 2008, 2008.
- [59] Mukta Patkar, Pooja Pawar, Mony Singh, and Ashwini Save. A new way for semi supervised learning based on data mining for product reviews. In *Engineering and Technology (ICETECH), 2016 IEEE International Conference on*, pages 819–824. IEEE, 2016.
- [60] Isidro Penalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia, Miguel Angel Rodriguez-Garcia, Valentin Moreno, Anabel Fra-
ga, and Jose Luis Sanchez-Cervantes. Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995–6008, 2014.
- [61] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.

- [62] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [63] Vidisha M Pradhan, Jay Vala, and Prem Balani. A survey on sentiment analysis algorithms for opinion mining. *International Journal of Computer Applications*, 133(9), 2016.
- [64] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
- [65] Reddit. <https://http://reddit.com/>.
- [66] Radim Rehurek and Petr Sojka. Gensima statistical semantics in python. 2011.
- [67] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. Interactive sankey diagrams. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 233–240. IEEE, 2005.
- [68] Mantosh Kumar Sarkar and Goutam Chakraborty. Opinion mining and geo-positioning of textual feedback from professional drivers. In *SAS Global Forum 2013 Proceedings*. Citeseer, 2013.
- [69] Mithileysh Sathiyaranarayanan and Donato Pirozzi. Spherule diagrams with graph for social network visualization. In *Communication Systems and Networks (COMSNETS), 2016 8th International Conference on*, pages 1–6. IEEE, 2016.
- [70] Azra Shamim, Vimala Balakrishnan, and Muhammad Tahir. Evaluation of opinion visualization techniques. *Information visualization*, 14(4):339–358, 2015.
- [71] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.
- [72] Prashast Kumar Singh, Arjit Sachdeva, Dhruv Mahajan, Nishtha Pande, and Amit Sharma. An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-*, pages 329–335. IEEE, 2014.
- [73] Paweł Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479, 2012.

- [74] Gamgarn Somprasertsri and Pattarachai Lalitrojwong. Extracting product features and opinions from product reviews using dependency analysis. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 5, pages 2358–2362. IEEE, 2010.
- [75] Zhiwen Song and Jianhong Cecilia Xia. Spatial and temporal sentiment analysis of twitter data. *European Handbook of Crowdsourced Geographic Information*, 205, 2016.
- [76] Stanford NLP Group. <https://nlp.stanford.edu/software/lex-parser.shtml>, 1990.
- [77] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25, 2017.
- [78] Maite Taboada, Julian Brooke, Milan Tofloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
- [79] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.
- [80] Pratik Thakor and Sreela Sasi. Ontology-based sentiment analysis process for social media content. *Procedia Computer Science*, 53:199–207, 2015.
- [81] Tom Anderson, Chris DeWolfe. <https://myspace.com/>, 2003.
- [82] Tomas Hurka, Jiri Sedlacek,. <https://visualvm.github.io/>, 2016.
- [83] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.
- [84] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. SentiView: Sentiment analysis and visualization for internet popular topics. *IEEE transactions on human-machine systems*, 43(6):620–630, 2013.
- [85] Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. Sentiment classification: The contribution of ensemble learning. *Decision support systems*, 57:77–93, 2014.

- [86] Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, and Daniel A Keim. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *VISW*, 2009.
- [87] Suchita V Wawre and Sachin N Deshmukh. Sentiment classification using machine learning techniques.
- [88] Wei Wei and Jon Atle Gulla. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 404–413. Association for Computational Linguistics, 2010.
- [89] World Bank. <http://www.worldbank.org/en/news/press-release/2014/08/28/world-bank-report-digital-payments-economic-growth>, 08 de agosto de 2014.
- [90] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1763–1772, 2014.
- [91] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–1118, 2010.
- [92] Kaiquan Xu, Stephen Shaoyi Liao, Jieyun Li, and Yuxia Song. Mining comparative opinions from customer reviews for competitive intelligence. *Decision support systems*, 50(4):743–754, 2011.
- [93] Zhijun Yan, Meiming Xing, Dongsong Zhang, and Baizhang Ma. Exprs: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, 52(7):850–858, 2015.
- [94] Christopher C Yang and Tobun Dorbin Ng. Analyzing and visualizing web opinion development and social interactions with density-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6):1144–1155, 2011.
- [95] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Ni black. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.

- [96] Jeonghee Yi and Wayne Niblack. Sentiment mining in webfountain. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 1073–1083. IEEE, 2005.
- [97] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652, 2008.
- [98] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 347–354. ACM, 2011.
- [99] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Constrained lda for grouping product features in opinion mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 448–459. Springer, 2011.
- [100] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O’Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 1462–1470. Association for Computational Linguistics, 2010.
- [101] Lili Zhao and Chunping Li. Ontology based opinion mining for movie reviews. In *International Conference on Knowledge Science, Engineering and Management*, pages 204–214. Springer, 2009.
- [102] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pages 338–349. Springer, 2011.
- [103] Lina Zhou and Pimwadee Chaovailit. Ontology-supported polarity mining. *Journal of the Association for Information Science and Technology*, 59(1):98–110, 2008.

