Análisis de los comentarios en español de usuarios de Facebook para la clasificación de publicaciones utilizando técnicas inteligentes

Tesina de Licenciatura en Informática

Alumnos: Gianetto, Emiliano Ariel
Saporiti, Lucía
Director: Dr. Hasperué, Waldo

1. **Introducción**
2. Redes Sociales
3. Procesamiento de Lenguaje Natural
4. Análisis de Sentimientos
5. Desarrollo propuesto
6. Estudio realizado
7. Conclusiones y trabajos futuros

Motivación

- **Redes sociales** son cada vez más utilizadas
- Poseen **millones** de usuarios día a día
- La interacción se traduce en **contenido**
- Mediante **técnicas** de Procesamiento de Lenguaje Natural y Minería de Datos se realiza el **Análisis de Sentimientos**
- Se analiza y obtiene **información** de interés
- **Actualmente** se investiga Análisis de Sentimientos
- La mayoría de los estudios son aplicados sobre textos en **inglés**
- Decidimos investigar y aplicar sobre textos en **español**

Objetivos

- Estudiar y aplicar **técnicas** de procesamiento de **lenguaje natural** para transformar los comentarios de una red social y luego analizarlos.
- Estudiar y evaluar la **performance** de las técnicas de clasificación estudiadas, planteando diversos escenarios de prueba.
- Implementar un **prototipo de aplicación** que, a partir de la opinión del usuario sobre comentarios, permita realizar un filtrado automático y muestre solo aquellos que puedan resultar de su interés.

1. Introducción
2. **Redes Sociales**
3. Procesamiento de Lenguaje Natural
4. Análisis de Sentimientos
5. Desarrollo propuesto
6. Estudio realizado
7. Conclusiones y trabajos futuros

¿Qué son las redes sociales?

- Conjunto de **grupos vinculados** unos a otros a través de relaciones sociales, y que tienen como fin la **interacción** de dos o más **actores**.
- En las **plataformas de Internet**, las personas interactúan a través de **perfiles** creados por ellos mismo, donde:
  - **comparten** sus fotos, videos, historias, eventos o pensamientos,
  - se **reúnen** para hablar, compartir ideas, hacer nuevos amigos y **socializar**,
  - **establecen** alguna relación y que mantienen **intereses** y **actividades** en común o se encuentran interesados en **explorar** los intereses y las actividades de otros usuarios.

Facebook

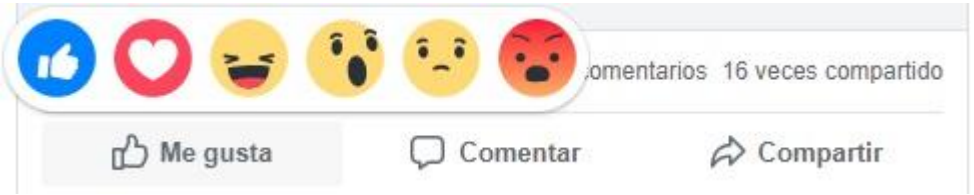
- Lanzado el 4 de febrero de **2004**. En español desde febrero de **2008**.
- Marzo de **2018**, contaba con más de **2200 millones de usuarios activos**.
- Inicialmente, para **estudiantes de Harvard.** Desde 2006, para cualquier persona mayor de **13 años**.
- **Nombre**: directorios de fotos personales que se entregan a estudiantes universitarios estadounidenses.
- **Acceder** desde una amplia gama de **dispositivos**.
- **Crear** un perfil personalizado: nombre, ocupación, escuelas, etcétera.
- **Utilizar** para asociar cuentas en otras redes sociales.

Servicios de Facebook

- **Grupos y páginas**: usuarios con **intereses comunes** o que puedan recurrir a ellos en **búsqueda** de algo puntual. Los **grupos** pueden ser públicos, cerrados o secretos y cualquier miembro puede añadir archivos. Las **páginas** son públicas y su administrador es el único que puede realizar publicaciones.

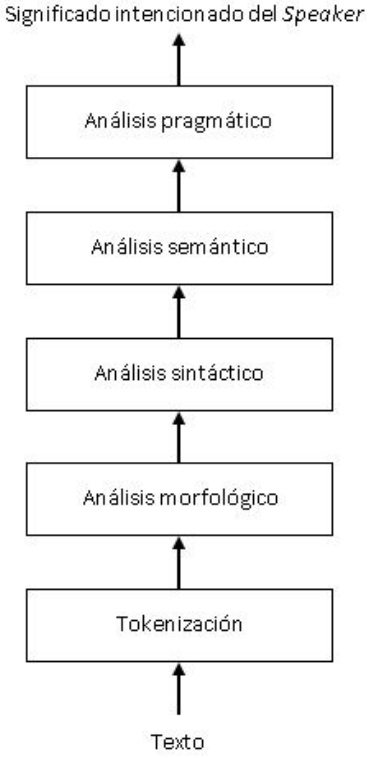
Servicios de Facebook

- **Botón «Me gusta»**: parte inferior de cada publicación o comentario hechos por los usuarios, **valorando su contenido**. También hay “**reacciones**”, permitiendo al usuario elegir el **nivel de su agrado** agregando las opciones “Me encanta”, “Me divierte”, “Me asombra”, “Me entristece” y “Me enoja”.



1. Introducción
2. Redes Sociales
3. **Procesamiento de Lenguaje Natural**
4. Análisis de Sentimientos
5. Desarrollo propuesto
6. Estudio realizado
7. Conclusiones y trabajos futuros

Procesamiento de Lenguaje Natural



- Ayuda a las computadoras a **comprender**, **interpretar** y **manipular** el lenguaje humano.
- **Comprensión** y **procesamiento** asistido de información para determinadas tareas.
- **Descompone** el proceso del análisis en varias **etapas**.
    - Diferencias entre ***sintaxis***, ***semántica*** y ***pragmática***.

Preprocesamiento de texto

- Mejora las **condiciones** del texto y la **efectividad** de los procesos futuros.
- Existen imperfecciones, errores, abreviaturas, jerga, o simplemente **datos que no nos interesan**.
- **Convertir** un texto sin procesar, en una **secuencia** bien definida de elementos (caracteres, palabras y oraciones), que serán utilizadas en etapas posteriores.

Tokenización

- **Convertir** una secuencia de caracteres en una secuencia de tokens.
- Métodos para identificarlos:
  - Expresiones regulares
  - Delimitadores (caracteres de separación)
  - Definición explícita por un diccionario

Técnicas de transformación de datos

- **Bolsa de palabras (BoW - Bag of Words)**: estructura de datos con las palabras y la cantidad de ocurrencias que tienen en un texto. Ignora la posición de la palabra en el documento.
- **Frecuencia de término - Frecuencia de documento inversa (TF-IDF - Term Frequency - Inverse Document Frequency)**: producto de la frecuencia de un término dentro de un texto (local) y qué tan específica es la palabra, tomando en cuenta el total de los textos (global).
- **Negaciones (Negations)**: analiza la presencia y alcance de la negación dentro de una frase, y evalúa cómo altera la opinión.
- **Reconocimiento de entidad con nombre (NER - Named Entity Recognition)**: etiqueta los elementos atómicos de la oración asignando categorías como: persona, lugar, entre otros.

Aplicaciones y ejemplos

- Descubrimiento de tópicos en colecciones de textos.
- Conversión de voz a texto, de texto a voz, y traducción automática.
- Resumen del documento.
- Correos electrónicos *spam*.
- **Análisis de sentimientos** (estado de ánimo, opinión).

1. Introducción
2. Redes Sociales
3. Procesamiento de Lenguaje Natural
4. **Análisis de Sentimientos**
5. Desarrollo propuesto
6. Estudio realizado
7. Conclusiones y trabajos futuros

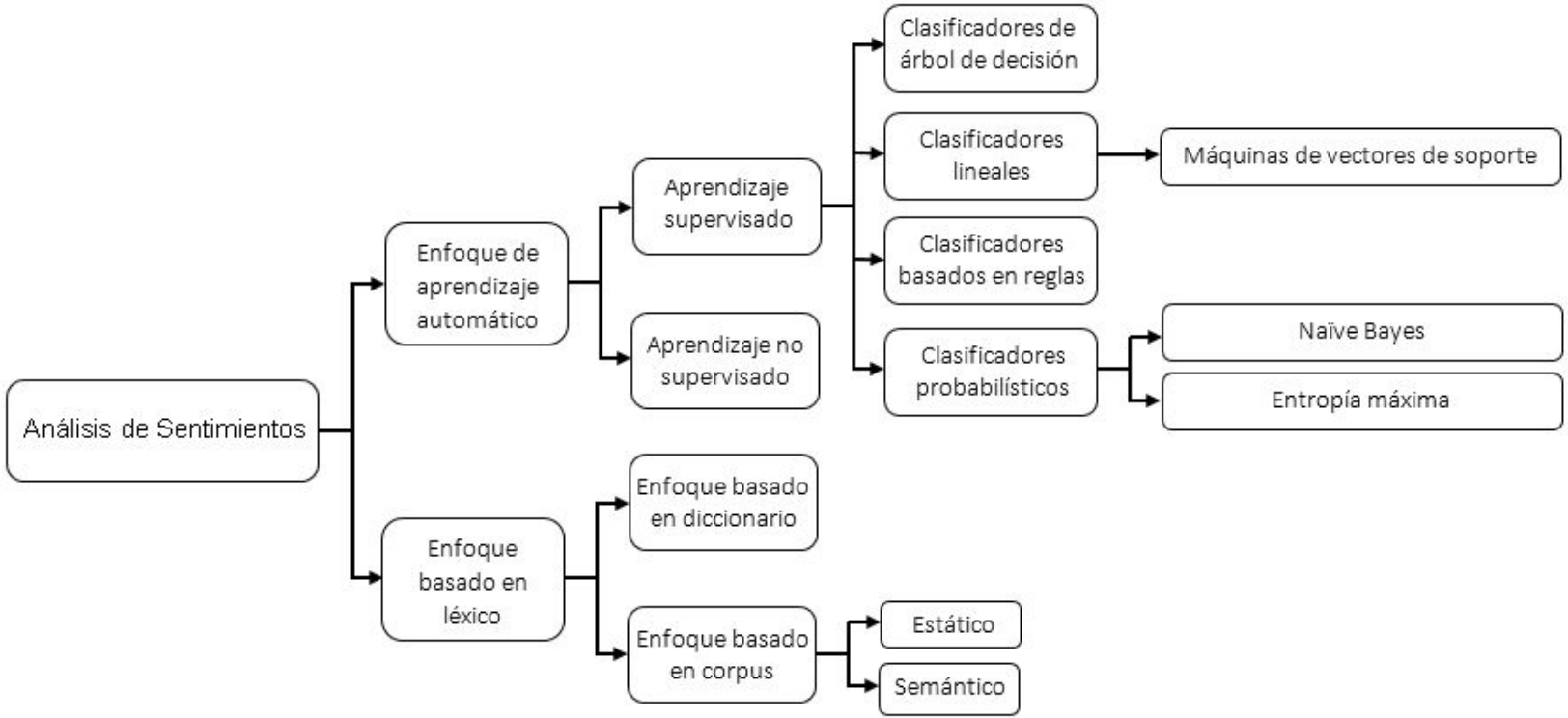
Minería de datos

- Gran cantidad de datos, de rápido crecimiento, recopilada y almacenada en grandes y numerosos repositorios de datos ➜ procedimiento **manual** es propenso a **errores**, **costoso** y **lento**.
- Necesidad de **obtener información útil** y conocimiento que pueda ser utilizado para la creación de diversas aplicaciones a partir del crecimiento exponencial que están sufriendo los datos.
- Resultado de la **evolución** natural de la tecnología de la información.

Análisis de Sentimientos

- **Estudio computacional** de opiniones, sentimientos, emociones y subjetividad expresadas en textos hacia una **entidad**.
- Su objetivo es **determinar** opiniones, **identificar** los sentimientos que expresa un escritor mediante su texto, y luego clasificar su **polaridad**.
- **Rapidez** y **espontaneidad** en que se obtiene la información ya que se realiza en el mismo momento en que se minan las opiniones.
- Realizar un seguimiento del **impacto**, comparando las métricas de **antes** y **después** del acontecimiento de un evento o noticia.
- **Dominios posibles**: productos, servicios, atención médica, artículos de noticias, eventos sociales, elecciones políticas, etcétera.

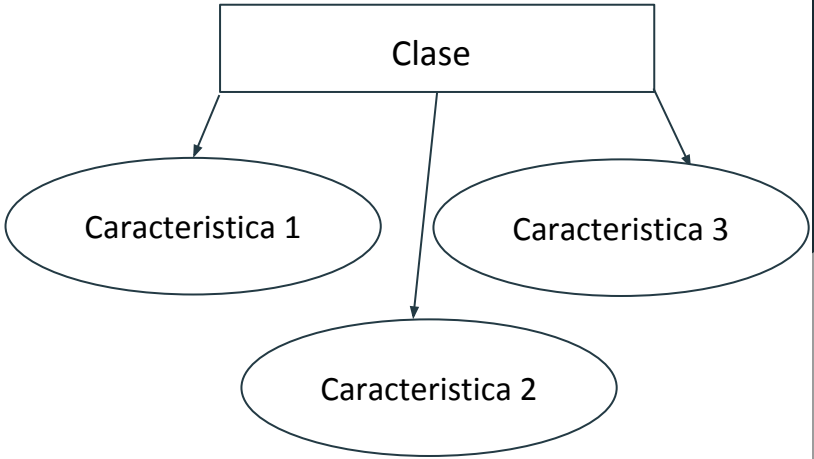
Técnicas para el Análisis de Sentimientos



Aprendizaje supervisado

- Clasificadores **probabilísticos** o **generativos**
  - Permiten la **descripción** del conjunto de observación a través de funciones de **probabilidad**.
  - **Describen** cómo se generan los datos.
  - Se pueden usar para imputar datos faltantes, comprimir el conjunto de datos o generar datos no vistos.
- Clasificadores **lineales** o **discriminativos**
  - **Identificar** a qué clase pertenece un objeto por medio de **funciones discriminativas** que mapean directamente la observación al valor de una etiqueta de clase.
  - **Superiores** en predicción, pero se obtiene **poco conocimiento** de los datos y cómo se generan

Naïve Bayes

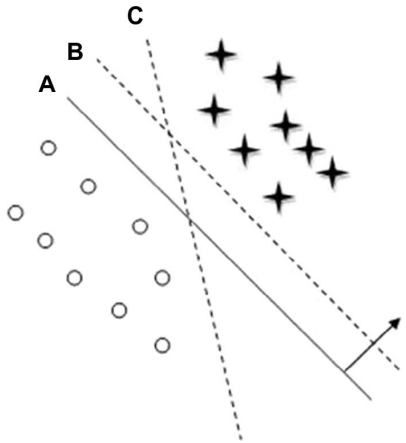


- Más simple, más utilizado, fácil de construir.
- Asume **independencia** entre características.
- Útil para **conjuntos** de datos muy **grandes**.
- Funciona bien en **múltiples clases**.
- Funciona mejor con variables de entrada nominales que con numéricas.
- Utiliza el **Teorema de Bayes** para predecir la probabilidad que un conjunto de características determinado pertenezca a una etiqueta en particular.

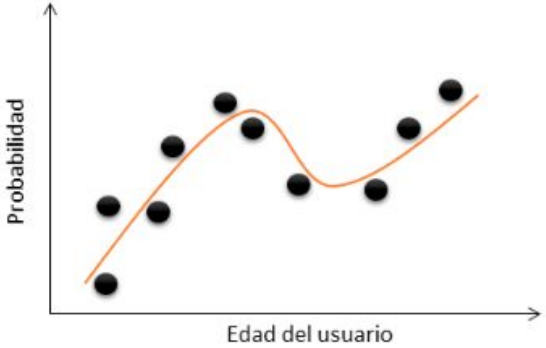
Entropía Máxima

- Pertenece a la clase de modelos exponenciales.
- No supone que las **características** sean condicionalmente independientes entre sí, toma en cuenta la **correlación** de las mismas.
- Predecir resultados posibles de una variable **distribuida categóricamente**.
- Estima las distribuciones de probabilidad de acierto a partir de datos usando **optimización iterativa**.
    - Comienza con **ponderaciones** previas **mínimas** y se optimiza para encontrar pesos que **maximicen** la probabilidad de los datos de entrenamiento.
- Aprendizaje es más lento que Naïve Bayes, y por lo tanto puede no ser apropiado dada una gran cantidad de clases para aprender.

Máquinas de Vectores de Soporte



- Representa los puntos de muestra en el espacio.
- Separa de forma óptima las clases mediante un hiperplano de separación al que se llama **vector soporte**.
- Los datos de texto son ideales, ya que pocas características son irrelevantes, pero tienden a correlacionarse entre sí.
- **Funciones Kernel**. Resuelven el problema de clasificación trasladando los datos a un espacio donde el hiperplano de solución es lineal. La técnica se entrena con una serie de datos de prueba.



Enfoque basado en léxico

Se basa en una una colección de términos de sentimiento conocidos y precompilados. Existen palabras que expresan una opinión **comparativa** sobre más de una entidad. Además hay expresiones o modismos, que forman un **léxico de opinión**.

- **Basado en el diccionario**: fácil y rápido para encontrar una gran cantidad de palabras de opinión y luego sus sinónimos y antónimos. Incapaz de encontrar palabras con el dominio específico del contexto.
- **Basado en corpus**: ayuda a resolver el problema de encontrar palabras de opinión específicas del contexto. Sus métodos dependen de patrones sintácticos o que ocurren junto con una lista de palabras de opinión.

Herramientas actuales

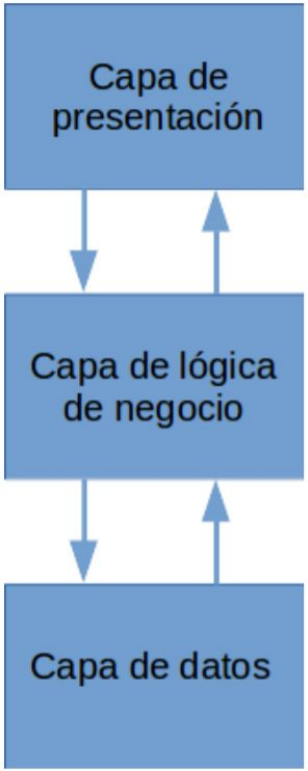
- **Text Analytics** (Microsoft): procesamiento de lenguaje natural avanzado sobre texto sin formato.
- **Google Cloud Platform** (Google) (Beta): servicios de aprendizaje automático con modelos ya preparados y está basado en redes neuronales.
- **Watson** (IBM): procesamiento del lenguaje natural, recuperación de información, representación del conocimiento, aprendizaje automático.
- Productos de caja negra, reciben datos y devuelven resultados.
- No es posible hacer uso del contexto.
- Google y Microsoft, no son gratuitos.
- Su uso no es amigable e intuitivo para el usuario final.

1. Introducción
2. Redes Sociales
3. Procesamiento de Lenguaje Natural
4. Análisis de Sentimientos
5. **Desarrollo propuesto**
6. Estudio realizado
7. Conclusiones y trabajos futuros

Desarrollo propuesto

- **Recolectar comentarios** conectándose con la API de Facebook y almacenarlos en una base de datos.
- **Etiquetar** un conjunto de comentarios de manera personal.
- **Analizar** el corpus, implementando un preprocesamiento del mismo.
- **Entrenar** y **aplicar modelos** basados en las técnicas de clasificación.
- **Evaluar** y **comparar** el **rendimiento** de los diversos modelos.

Arquitectura de la aplicación



Interfaz gráfica. Nivel más alto de la aplicación. Captura la interacción del usuario y muestra la información.

Funcionalidad. Procesa las peticiones de la capa superior y obtiene la información de la capa inferior.

Información. Almacena y recupera información de la base de datos.

Descripción y funcionamiento - Recolección

- Se comunica con API
- Recolecta datos
- Posee corte de control
- Almacena todo en la base de datos

Descripción y funcionamiento - Etiquetado

Cambio de rumbo

- Utilizar **reacciones** de usuarios sobre las publicaciones para **complementar** el análisis del texto de su **comentario.**
- **Cambios** en la política de **permisos** de la API de Facebook.
- Imposible obtener la información del **usuario** que generó cada **reacción**.
- Imposible obtener **reacciones** a los **comentarios**.
- Conjunto de comentarios etiquetados **manualmente**.
- Resultados **dependen** de la opinión propia de los tesistas.
- Aplicación **personalizable.**

Descripción y funcionamiento - Análisis

1. Decodificación de formato utf-8 a latin-1
   “MaÃ±ana se levantarÃ¡ la sesiÃ³n? JaJaJa” -> “Mañana se levantará la sesión? JaJaJa”
2. Pasaje a minúsculas
   “Mañana se levantará la sesión? JaJaJa” -> “mañana se levantará la sesión? jajaja”
3. Separación de palabras
   “mañana se levantará la sesión? Jajaja” -> [“mañana”, “se”, “levantará”, “la”, “sesión”, “jajaja”]
4. Normalización de risas
   “jajaja” -> “inter-risa”
5. Eliminación de stopwords (singular y plural)
   [“mañana”, “se”, “levantará”, “la”, sesión”, “jajaja”] ->
   [“mañana”, “levantará”, “sesión”, “inter-risa”]

Descripción y funcionamiento - Entrenamiento

Lexicon:

- 12000(inicial) + 500 palabras
- Polaridad = promedio de valores

Machine Learning:

- Naïve Bayes, ME y SVM
- Validación cruzada de K iteraciones
- Bag of Words

1. Introducción
2. Redes Sociales
3. Procesamiento de Lenguaje Natural
4. Análisis de Sentimientos
5. Desarrollo propuesto
6. **Estudio realizado**
7. Conclusiones y trabajos futuros

Caso de estudio

- **Red social:** Facebook
- **Portal:** Todo Noticias
- **Publicaciones:** 569
- **Comentarios:** 270279
- **Etiquetados:** 3000
- **Parámetros modificados:**
  - Cantidad mínima de tokens
  - Número de iteraciones
  - Tamaño del BoW.

Etapa de entrenamiento

Lexicon:

- Intervalo [-1, 1] y límites L1 y L2.
- -1 < L1 < L2 < 1
- Límites iniciales: L1=-0.3 y L2=0.3
- Límites finales: L1=-0.009 y L2=0.05

Machine Learning:

- Mínimo de 1 (todos) y 2 tokens
- Validación cruzada de 3, 5 y 10 iter.
- Tamaño del BoW de 73 y 108 tokens

Tokens	Pos	Neu	Neg
>=1	1000	1000	1000
>=2	884	865	889
>=3	721	661	768
>=4	604	420	652

Evaluación de técnicas

Lexicon:

- **Ensayo 1:** Sin filtrar stopwords. Límites: -0.3, 0.3.
- **Ensayo 2:** Sin filtrar stopwords. Límites. -0.009, 0.05.
- **Ensayo 3:** Filtrando stopwords. Límites: -0.3, 0.3.
- **Ensayo 4:** Filtrando stopwords. Límites: -0.009, 0.05.

Evaluación de técnicas

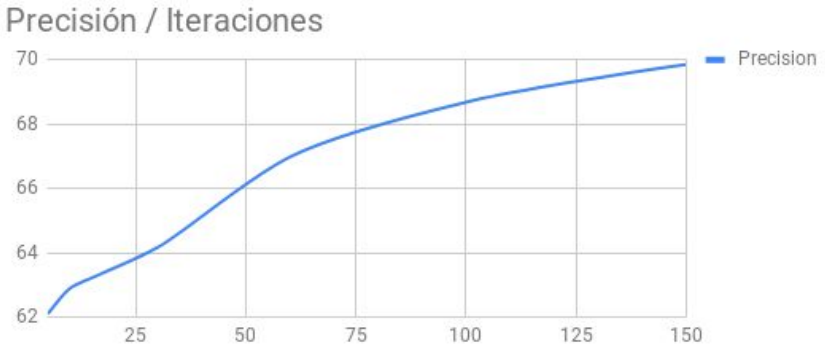
Machine Learning:

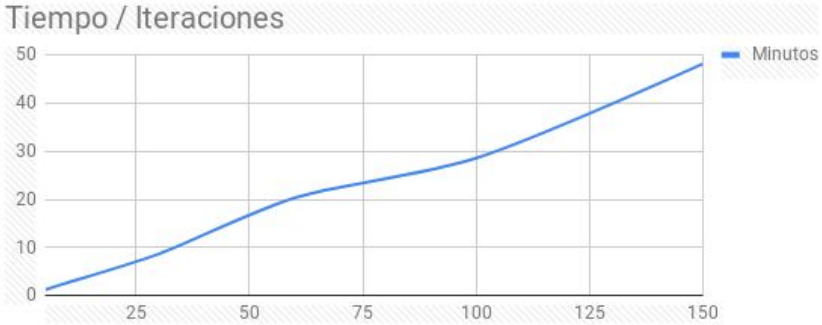
Ensayo	Tokens	Folds	Tamaño BoW
1	>=2	3	108
2	>=2	5	108
3	>=2	10	108
4	>=2	3	73
5	>=2	5	73
6	>=2	10	73

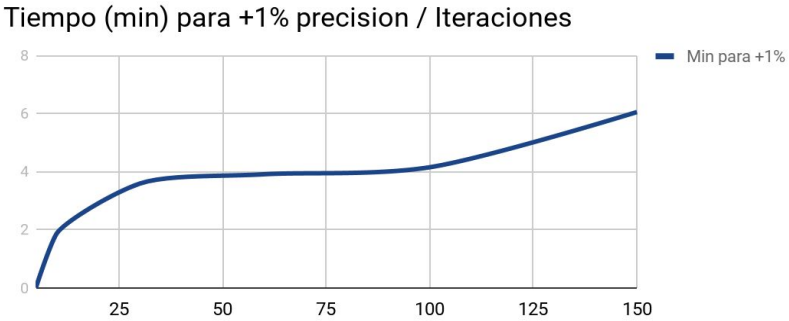
Ensayo	Tokens	Folds	Tamaño BoW
7	>=1	3	108
8	>=1	5	108
9	>=1	10	108
10	>=1	3	73
11	>=1	5	73
12	>=1	10	73

Evaluación de técnicas - ME

Iters	Minutos	Precisión	Prec. extra	Min para +1%
5	1.28	62.10%	-	-
10	2.77	62.90%	+0.80%	1.86
30	8.70	64.17%	+2.07%	3.58
60	20.35	66.98%	+4.88%	3.91
100	28.58	68.68%	+6.58%	4.15
150	48.18	69.85%	+7.75%	6.05



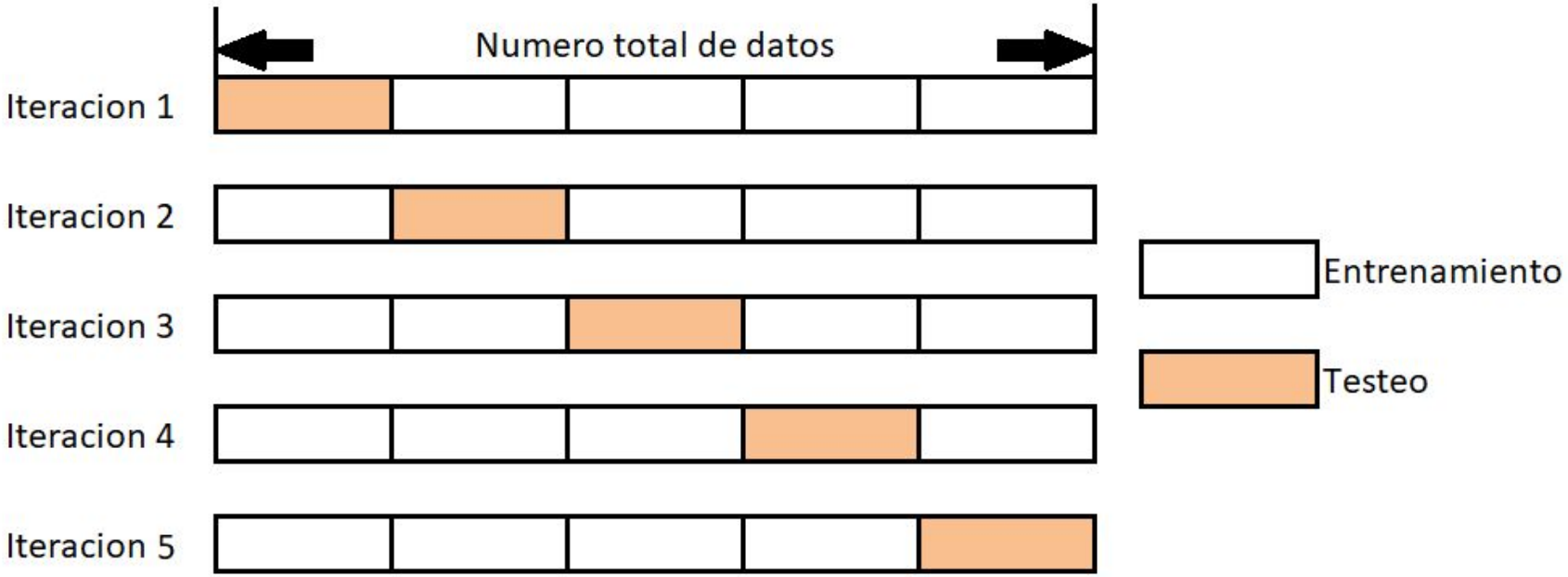




Performance

- **Análisis de desempeño** o *performance* del modelo, para saber qué tan exitoso resultó el entrenamiento.
- Existen diferentes **métricas** utilizadas para evaluar los algoritmos de aprendizaje automático.
- Obtuvimos solo aquellas utilizadas para los problemas de clasificación.

Validación cruzada de K iteraciones



Matriz de confusión

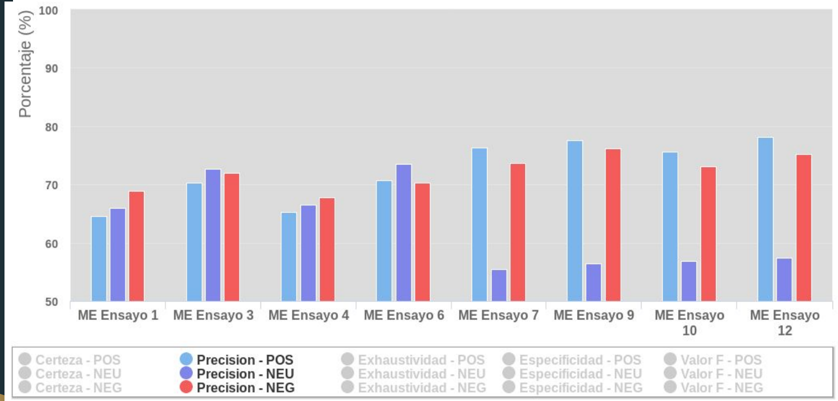
		Real		
		Pos	Neg	Total
Predicción	Pos	**40**	**20**	60
	Neg	**30**	**10**	40
	Total	70	30	

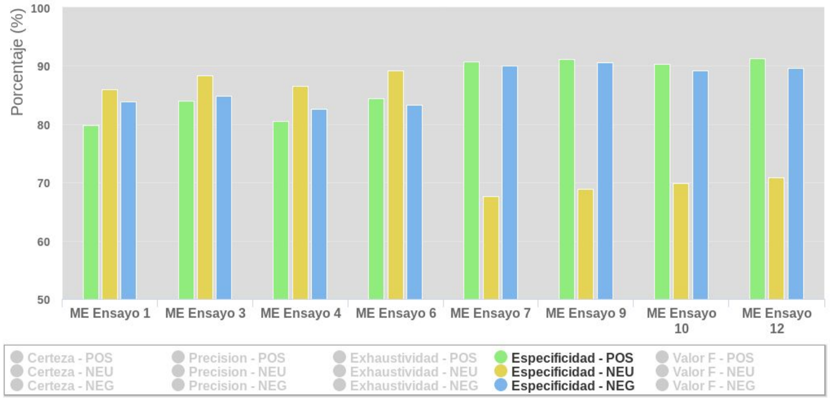
Métricas

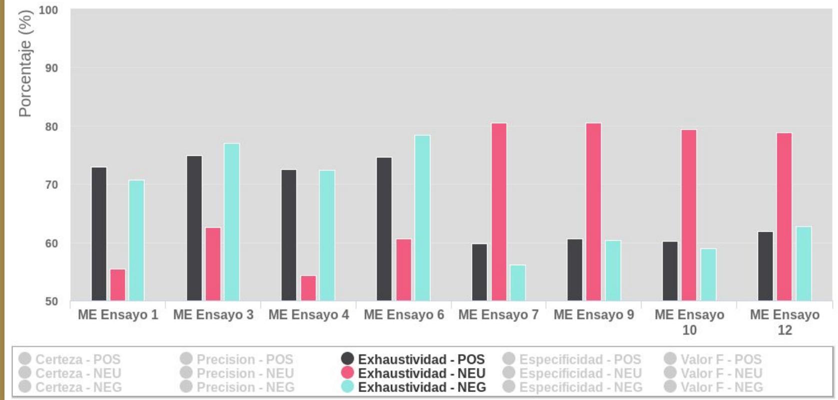
- **Certeza (accuracy):** predicciones correctas respecto del total de predicciones.
- **Precisión (precision):** predicciones correctas respecto de las predicciones para una clase.
- **Exhaustividad (recall):** casos positivos cuya predicción fue acertada.
- **Especificidad (specificity):** casos que no pertenecen a la clase *X*, y fueron correctamente predichos como otra clase.
- **Valor-F:** promedio armónico entre Precisión y Exhaustividad.

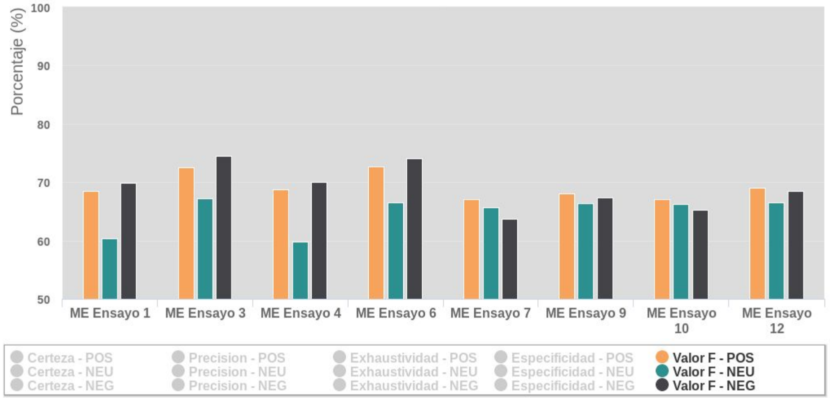
		Real	
		Pos	Neg
Predicción	Pos	**40**	**20**
	Neg	**30**	**10**

Comparación de métricas para las técnicas (ME)

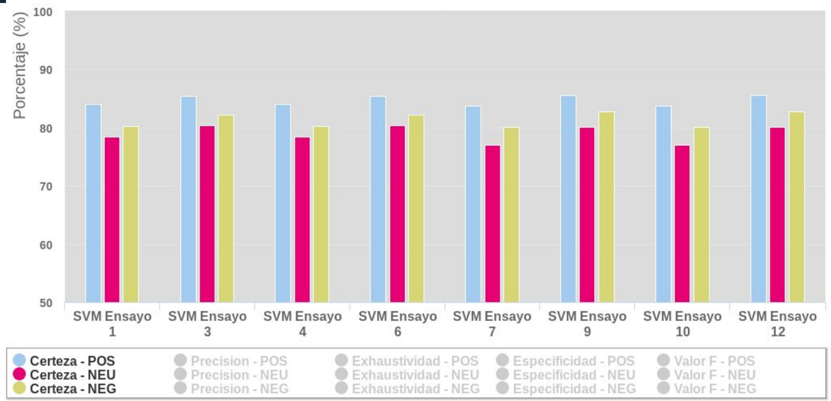


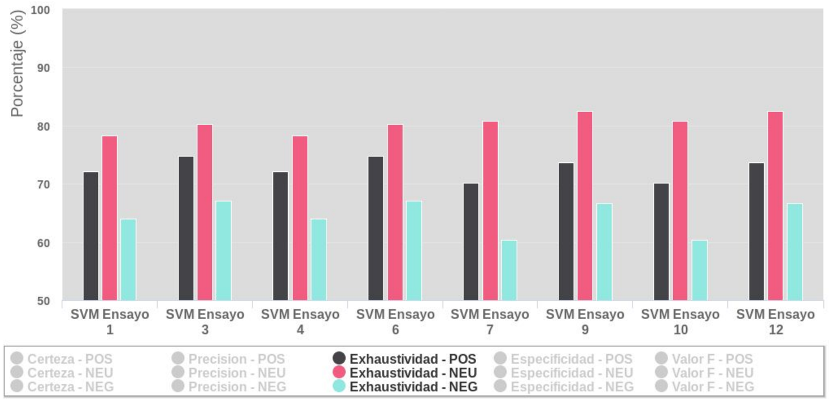


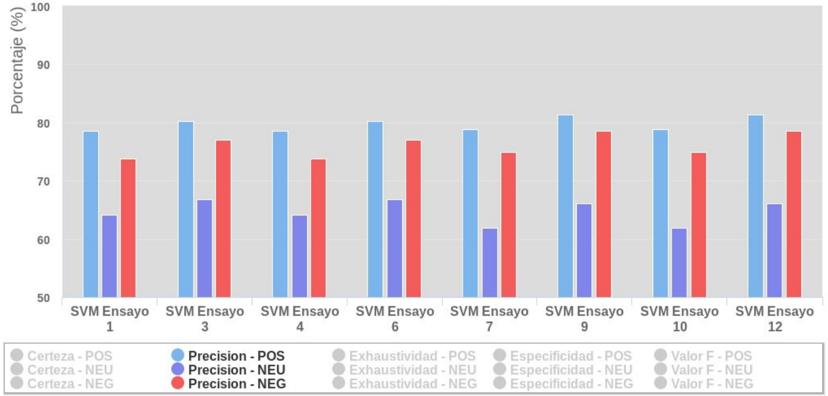


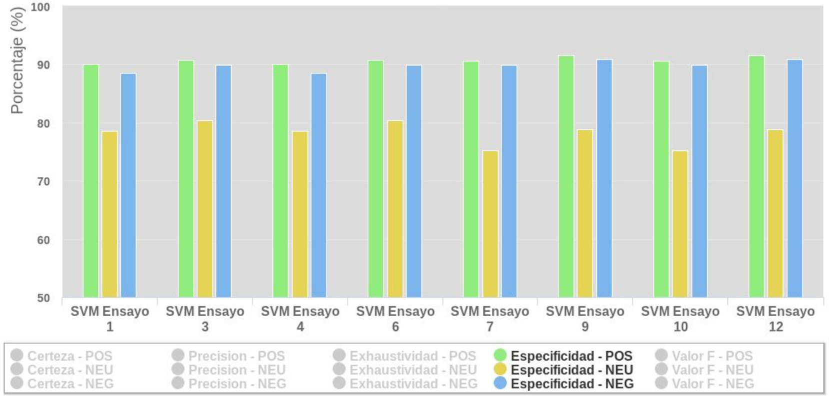


Comparación de métricas para las técnicas (SVM)

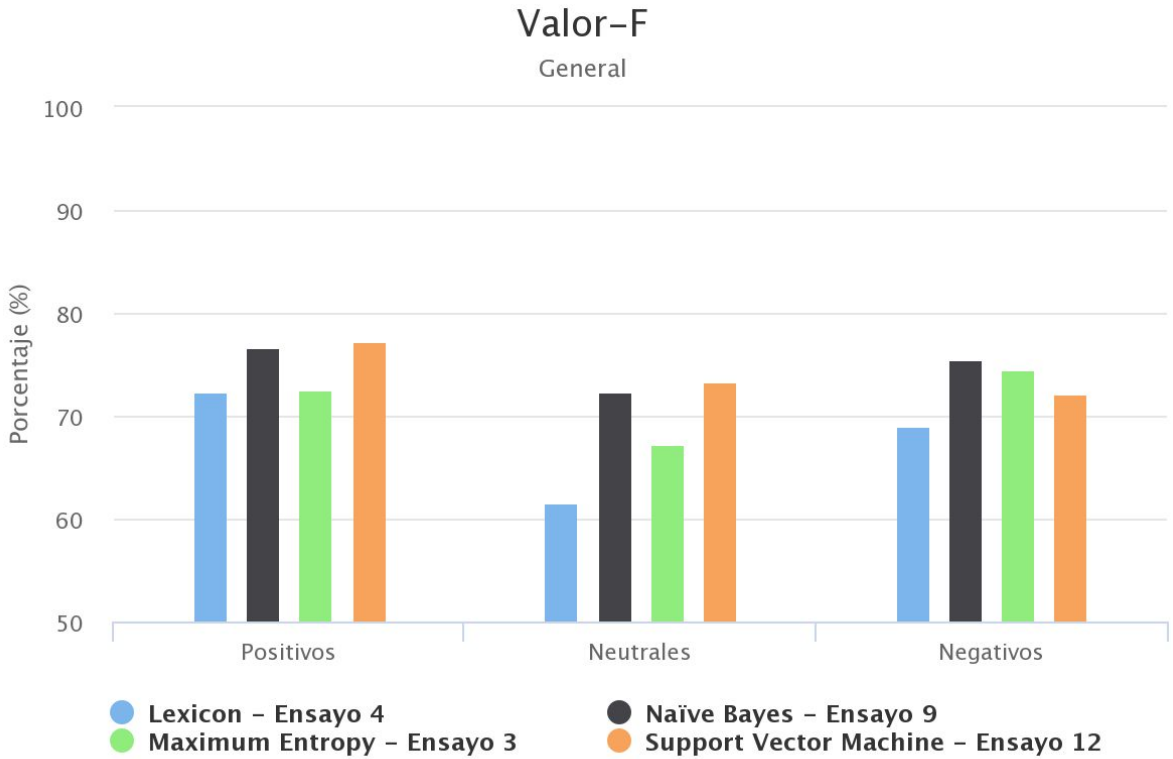








Comparación de métricas para las técnicas



1. Introducción
2. Redes Sociales
3. Procesamiento de Lenguaje Natural
4. Análisis de Sentimientos
5. Desarrollo propuesto
6. Estudio realizado
7. **Conclusiones y trabajos futuros**

Conclusiones

- Los **límites** del **Lexicon** tienen gran repercusión en la performance.
  - Filtrar las **stopwords** mejora el desempeño.
- **Incrementar** el número de **folds** mejora los resultados en las técnicas de aprendizaje automático.
  - **NB** mejora al utilizar un **mayor** tamaño de **BoW.**
  - **MaxEnt** posee **dificultades** para identificar la clase neutral.
  - **SVM** mejora levemente al utilizar comentarios más **extensos** y **menor** tamaño de BoW.
- Los **mejores** resultados fueron obtenidos por **NB** y **SVM**.

Trabajos futuros

- **Influencia** entre **usuarios** dentro de un hilo de discusión. **Variación** de **opinión**.
- **Agrupar usuarios** en base a la **similitud** de su **valoración** sobre los noticias/publicaciones.
- Extraer **información** sobre el **usuario. Ranking** de noticias/publicaciones más aceptadas o rechazadas para **grupos** con ciertas **características similares**.
- Permitir que el **usuario** pueda **reaccionar**, de forma similar a Facebook, sobre los comentarios en **tiempo real**.
- **Recolectar** publicaciones y comentarios según lo **aprendido** por el **modelo**.
- Posibilitar la **descarga** de una base o "**perfil de opinión**" de otra persona, para **evitar** la necesidad de **etiquetar** comentarios antes de poder utilizar la aplicación.

¿Preguntas?