[[1]](#footnote-1)

Diseño e Implementación de un Clasificador de Sentimientos

Stalyn Quishpe, Andres Paredes

*Abstract*— in this document it talk about sentiment analysis, your definitions in technology information. We present a lexicon-based approach to extracting sentiment from text the book of project Gutenberg and determine if it prayers are positive, negative or neutral. It uses dictionaries of words that they are classified in verbs and adjectives, also we going to create a model based on this classification to predict if a sentence is positive, negative or neutral in base of an algorithm of learning.

*Index Terms*—Sentimental analysis, project Gutenberg, prayers positive, prayers negative.

# Introducción

E

stamos en la época donde la tecnología se encuentra en cualquier escenario como son; empresas, instituciones dedicadas a la educación, servicios de transporte, banca, etc. La información que producen estas instituciones es de gran ayuda para la toma de decisiones y así poder mejorar la lógica, el manejo y las ganancias del negocio. La correcta manipulación de esta información involucra varios aspectos tanto tecnológicos como sociales, políticos, educativos, sentimentales, etc.

Una característica importante del comportamiento humano es la recopilación de información, con esta información y un análisis adecuado se tiende a clasificación de la misma para facilitar su uso, de la misma manera posteriormente se tiende a la predicción de acciones y demás. En la actualidad se cuenta con un gran número de fuentes de datos para la recopilación de información como son redes sociales, blog personales, sitios en línea entre otras, se puede utilizar estas fuentes de datos para buscar, almacenar y entender opiniones de los demás acerca de algún tema en concreto, para posteriormente dales un uso a los datos obtenidos por este análisis.

Un problema actual el cual está teniendo una gran acogida es el “Análisis de Sentimientos”, puesto que este sirve para para identificar y extraer información subjetiva de ciertos recursos. El problema en el análisis de sentimientos es lo complejo y estructurado del lenguaje natural y el planteamiento de una solución robusta y precisa para desarrollar un técnica para el análisis de sentimientos. Este inconveniente también se debe a que las oraciones son subjetivas y cada opinión puede ser diferente, debido a una gran colección de opiniones en la web y las diferentes jergas que podemos encontrar en internet y emoticones cada día se torna más difícil analizar sentimientos.

Este documento se centra en el análisis de sentimientos utilizando un programa desarrollado por los autores de este artículo, y con la ayuda de herramientas para el aprendizaje automático. El análisis de sentimientos se aplicara a un libro obtenido en formato .txt del repositorio proyecto Gutenberg, Se lo analiza mediante una orientación semántica que es un medida de subjetividad y de opinión en el texto, también se evaluara mediante un factor de evaluación (positivos, negativo y neutro) hacia un tema, persona o idea.

Existe un enfoque principal para el problema de analizar y extraer sentimientos de una manera automática [1].

1. El enfoque basado en el léxico:

* Este fomenta el desarrollo de la capacidad lingüística de la persona mediante el aprendizaje de bloques prefabricados de palabras (chunks).
* En este caso se presenta un grupo de chunks de adjetivos y verbos clasificados en positivos y negativos.
* Esto entra en la definición “categorías de aspectos y expresiones de aspectos”.
* Una categoría de aspectos representa un único aspecto de la entidad por ejemplo: adjetivos o verbos.
* Una expresión de aspectos es una palabra real o frase que aparece en el texto que indica una categoría del aspecto por ejemplo: good, like [2].

# METODO

Como se mencionó en la introducción el problema que se encuentra en el análisis de sentimientos es la dificultad de concluir si nuestro texto expresa algo positivo, negativo, neutro. Para explicar cómo se logró solucionar este problema se partirá de un ejemplo que es un comentario sobre una cámara Canon [3].

Posteado por: John Smith Date: Septiembre 10,2011

“(1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) However, my wife thinks it is to heavy her”

A partir de este post vamos a realizar el proceso de análisis de sentimientos:

1. Procedemos a limpiar el texto quitando ciertos caracteres especiales y eliminando espacios en blanco. También cada oración se divide por un punto (“.”).

* En el ejemplo planteado observamos que divide en 5 oraciones.
* Estas oraciones a la vez las dividimos en palabras.

1. Una opinión se estructura de dos componentes: un target “g” y un target valor de sentimiento “s”
   1. “g”: Es una de las palabras que están definidas en el chunks en este caso trabajamos con chunks de adjetivos y verbos. Estas tienen También puede tomar un valor que está definido en una escala 1 al 5.

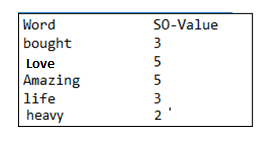


Fig. 1: Ejemplo de polarización de las palabras que se encuentran en el chunk de adjetivos y verbos.

* 1. “s”: es el sentimiento positivo, negativo o neutro.

1. Por ejemplo: si analizar la primera oración (1) y la última oración (5).

* I **bought** a Canon G12 camera six months ago.

Analizamos esta oración y en la búsqueda que realiza nuestro programa encuentra la palabra “**bought**” y le asigna un valor de 3, como ya las palabras que siguen no se encuentran en el chunk, nos da como resultado 3. Por lo tanto deducimos que nuestra oración al no tener más palabras que pertenezcan a un chunk y la sumatoria de las polaridades de las palabras es positiva por lo tanto esta se clasificara como positiva.

* However, my wife thinks it is to **heavy** her

Analizamos la siguiente oración y en la búsqueda que realiza nuestro programa encuentra la palabra **“heavy”** y le asigna un valor de 2, como ya las palabras que siguen no se encuentran en el chunks, nos da como resultado 2. Al ser este clasificado como una palabra negativa en el procesamiento tomara un valor negativo y en ausencia de otras palabras que se puedan analizar, procedemos a determinar que nuestra oración es negativa.

Si analizamos todo el comentario sobre la cámara Canon observamos que existen dos opiniones. La opinión del que muestra conformidad por la cámara y la opinión de vista que da una crítica a la cámara. Pero como existen más oraciones positivas que negativas en el comentario si puede concluir que el comentario de la cámara es positivo. Sin embargo para nuestro propósito nuestro procesamiento solo determinara la polaridad de las oraciones.

Para poder analizar estar opiniones se utilizó el concepto de opiniones regulares que a su vez está clasificada en dos: Opinión directa y opinión indirecta.

Una opinión directa: Es la más simple de analizar y esta es una opinión que expresa un aspecto en concreto de una entidad. Por ejemplo: “The picture quality is great” [4].

Por otra parte la escala o el valor que posee cada palabra en cada chunk esta evaluada según la gravedad de la palabra en sí, esto se debe a que en el lenguaje existen ciertas palabras que tienen un peso mayor que otras en las expresiones, cabe mencionar que solo las palabras más relevantes tienen valores superiores a 1 caso contrario por defecto a cada palabra se le otorga un valor de 1 si es un adjetivo y 2 si es un verbo.

La clasificación realizada en el proceso mencionado anteriormente se aplicara sobre un texto que será usado como un training data set para la generación de un modelo de predicción dicho modelo será procesado con un algoritmo Naïve Bayes y J48. Todo esto con el objetivo de crear un modelo capaz de predecir la polaridad que posee una oración es decir un modelo que pueda clasificar un una oración en buena, mala o neutral.

Por una parte el algoritmo Naïve Bayes es uno de los clasificadores más utilizados gracias a su simplicidad y rapidez.

Esta es una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados. Constituye una técnica supervisada puesto que se necesita tener ejemplos clasificados para que funcione.

Por otro lado J48 es una implementación del algoritmo C4.5 propia de la herramienta Weka este genera un árbol de decisión en base a un grupo de datos de entrenamiento, y permite trabajar con valores continuos ademas permite el manejo de datos de entrenamiento con datos faltantes.

Para la creación del modelo procederemos a la creación de un archivo para realizar el respectivo entrenamiento, dicho archivo será generado a partir del proceso de análisis de sentimientos establecido anteriormente basado en el concepto de opiniones regulares directas. Para lo cual se implementó un programa en lenguaje Pearl que se adjuntara como anexo el cual realiza el proceso previamente explicado y generara un archivo. Arff el cual procederemos a usar para la creación de nuestro respectivo modelo. Para esto se usara la app de la herramienta Weka especializada para lenguaje java. Esto facilitara la creación y evaluación del modelo mediante una validación cruzada.

“Validación Cruzada” es una técnica utilizada que permite evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba [5].

A continuación se procederá a usar el modelo creado con un conjunto de datos sin clasificar y se probara que tanta precisión tiene nuestro modelo usado con datos nuevos y se presentara una matriz de confusión para mostrar las instancias clasificadas correctamente e incorrectamente respectivamente.

# Resultados

Los siguientes resultados fueron producidos mediante la app de weka para java.

1. Naive Bayes

Los primeros resultados a mostrar corresponden a un modelo creado con el algoritmo Naive Bayes.

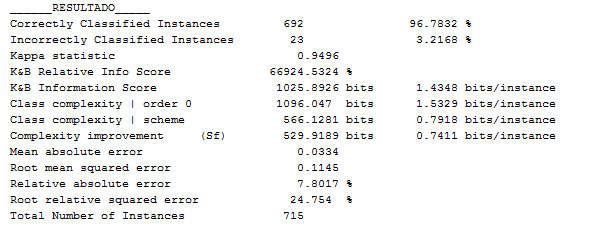


Fig.2 Resultados obtenidos por Naive Bayes

Aquí se muestran la cantidad de instancias clasificadas correcta e incorrectamente por el modelo, así mismo se muestra el error absoluto y relativo obtenido. Como se puede observar el porcentaje de instancias clasificadas correctamente muestra un valor de 96% y un error absoluto de 7%.

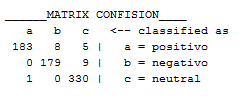


Fig.3. Matriz de confusión obtenida

La matriz de confusión nos muestra la cantidad de falsos positivos y resultados clasificados correctamente por cada clase existente. Como podemos ver el número de instancias clasificadas como falsas positivas o mal clasificadas es un valor relativamente bajo con respecto al de las instancias clasificadas correctamente, esto se repite en cada una de las clases que se tiene.

1. J48

El siguiente segmento del documento nos muestra los resultados obtenidos mediante el uso del algoritmo J48 (C4.5)

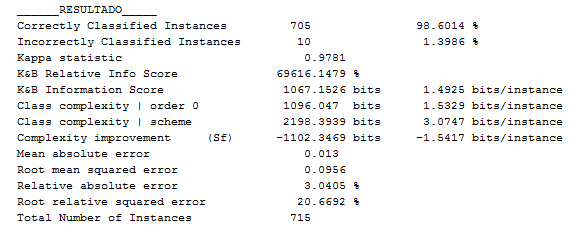


Fig.4. Resultados obtenidos por algoritmo J48

La figura 4 nos muestra que el número de instancias clasificadas correctamente por el algoritmo J48 es de 98%, mientras que el error absoluto generado es del 3%.

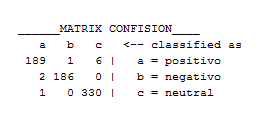


Fig.5. matriz de confusion

Por otro lado la matriz de confusión nos ofrece resultados similares al anterior, al referirnos al número de instancias clasificadas correcta e incorrectamente en cada clase.

# Discusión

Al comparar los resultados obtenidos por cada método podemos observar que cada uno de estos tiene un porcentaje alto de precisión al clasificar instancias, sin embargo el error absoluto que se genera por cada modelo es diferente, por un lado con un modelo generado por Naive Bayes nos da un error de 7%, el cual comparado con el 3% obtenido por el algoritmo J48 es más del doble. Esto se puede comprobar por los resultados obtenidos por en las respectivas matrices de confusión, aunque de igual manera muestras resultados relativamente positivos con respecto a instancias clasificadas correctamente, el algoritmo J48 muestra una cantidad menor de instancias clasificadas incorrectamente.

Ahora debemos considerar que el modelo se puede considerar sobre entrenado, puesto que en el archivo con el que se realiza el entrenamiento consta de un numero de instancias que se encuentran repetidas, esto se debe los atributos que se toma en cuenta para realizar la clasificación inicial y generar el archivo .arff son numéricos lo que provoca que los atributos se repitan en muchas ocasiones y el modelo se especialice para solo este conjunto de datos y provoca que no se produzca una clasificación optima con nuevos valores.

# Conclusión

## El análisis de sentimientos u “opinión minig” se puede considerar como un trabajo de alta complejidad puesto que para analizar y obtener resultados 100% correctos se debe considerar que existe una gran subjetividad en las opiniones de las personas, también por el hecho que el lenguaje natural es muy complejo y existen varias variantes de una palabras pero continúan significando lo mismo por ejemplo: Moto, motora, etc.

1. Se establece que los conceptos de subjetividad y emoción son altamente relacionados pero estos no tiene ninguna relación con la opinión. Pero se debe tomar en cuenta que no existe un conjunto claro de emociones existentes tabuladas o formalizadas. Por lo tanto el análisis de sentimiento está basado en principalmente en la evaluación de opiniones que expresen algo positivo o negativo.
2. El modelo se considerar sobre entrenado por la cantidad y subjetividad de los atributos tomados para realizar la clasificación de las oraciones, esto genera que instancias diferentes generen los mismos atributos. Esto se debe a que los atributos son generados en base a un proceso de polarización de las oraciones usando opiniones directas sin considerar opiniones indirectas como se explicó anteriormente en este documento.
3. Analizando los resultados obtenidos por el clasificador de Bayes tienen un mayor error respecto al J48 (c4.5) sin embargo este es de gran ayuda para prácticas donde los atributos son totalmente independientes entre sí y estos son numéricos, de esta manera se puede generar un modelo suficientemente robusto para algunos caso donde se puede ignorar deficiencias. Cabe mencionar que en nuestro caso de estudio los atributos que se toman para la clasificación no son totalmente independientes entre sí.

# Trabajos Futuros

Para trabajos futuros se implementará un analizador de sentimiento para el lenguaje Español (castellano) en un lenguaje de programación de más bajo nivel como C, para así lograr que el procesamiento del texto sea mucho más rápido y efectivo. También se aplicaran nuevas estrategias para aumentar la precisión de nuestro analizador, tomando en cuenta los valores obtenidos en este presente trabajo.

References

1. Devitt, Ann and Khurshid Ahmad. 2007.Sentiment polarity identification in financial news: A cohesion-based approach. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 984–991, Prague.
2. Ding, Xiaowen and Bing Liu. Resolving Object and Attribute Coreference in Opinion Mining. in Proceedings of International Conference on Computational Linguistics (COLING-2010). 2010.
3. Abdul-Mageed, Muhammad, Mona T. Diab, and Mohammed Korayem. Subjectivity and sentiment analysis of modern standard Arabic. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers. 2011.
4. Pang, Bo and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008. 2(1-2): p. 1-135.
5. es.wikipedia.org. (17 de 08 de 2016). Obtenido de https://es.wikipedia.org/wiki/Validaci%C3%B3n\_cruzada

1. [↑](#footnote-ref-1)