Análisis de Sentimientos
Un caso en reseñas de películas

Ybrahin Martínez, Josseline Perdomo

Escuela de Computación
Facultad de Ciencias
Universidad Central de Venezuela (UCV)
Av. Paseo Los Ilustres, Los Chaguaramos, Caracas, Venezuela
ybrahin.martinez,josseline.perdomo,(@ciens.ucv.ve)

Resumen

El Análisis de Sentimientos *es el estudio computacional de cómo las opiniones, actitudes, emociones y perspectivas son expresadas el lenguaje [2]. La competición* Bag of Words Meets Bags of Popcorn *del portal especializado en competencias de minería de datos* Kaggle *[1] propone un caso aplicado a los reseñas de películas del portal IMDb, proporcionando datos para resolver el problema usando* Aprendizaje Supervisado *o* No Nupervisado*, en el que cada reseña debe ser clasificado como positivo o negativo. Muchas de las soluciones publicadas por varios investigadores del área se enfocan en usar técnicas de aprendizaje supervisado como Naïve Bayes, Máxima Entropía o SVM a pesar de que existe una gran variedad de clasificadores actualmente. Este paper se basa en la solución propuesta por H. Parmar, S. Bhanderi y G. Shah aplicando* Random Forest *[3], con dos modalidades de generar el* Words Vector *generando interesantes resultados, con una precisión (accuracy) de 0.84, un valor bastante elevado en una aplicación de minería de datos en la que predomina la ambiüedad y los datos algunas veces son complicados de manejar.*

1. Introducción

El incremento de los datos por parte de opiniones en blogs, redes sociales, comercios electrónicos ha aumentado exponencialmente en la última década; esta gran cantidad de datos está jugando un papel muy importante en la toma de decisiones. La razón es clara, las organizaciones quieren analizar y evaluar el impacto de sus productos, desean conocer los gustos de sus consumidores, empleando esta técnica, como consecuencia, se ahorran una importante cantidad de tiempo y dinero que sería invertida en sondeos, obteniendo los mismos resultados. No sólo las organizaciones evalúan las opiniones en la internet, hoy en día la mayoría de las personas no compran cosas sin hacer antes un análisis de productos en internet, la gente busca comentarios que verifiquen su calidad y luego toma su decisión. En la actualidad, podría aseverarse que gran parte de la toma de decisiones es social, por lo que analizar los sentimientos juega un rol fundamental.

El análisis de sentimientos es un tema difícil en *Machine Learning*. Las personas expresan sus emociones en un lenguaje que a menudo es oscurecida por el sarcasmo, la ambigüedad y juegos de palabras, todo lo cual podría ser muy engañoso para los seres humanos, más aún para las computadoras. [1]. Existe una amplia gama de estados de ánimos y emociones que un humano puede expresar, sin embargo, éstos podrían agruparse en dos principales categorías: buenos y malos.

Este paper toma como caso de estudio las reseñas a películas proporcionadas por usuarios del portal IMDb [6], bajo un enfoque de *Aprendizaje Supervisado*, en el que cada una de las reseñas que forman parte del conjunto de entrenamiento es denominada como buena o mala, siguiendo la simplificación de sentimientos que hemos comentado anteriormente.

2. Análisis Exploratorio de Datos

El conjunto de datos está constituidos de tres características o columnas, el identificador unívoco de la reseña, el texto de la reseña en idioma inglés y un numero entre 1 o 0 que indica el sentimiento, bueno o malo respectivamente. Se vió como era la proporción de las reseñas y se encontró que estaban completamente balanceadas. La tabla de abajo se encuentran el número de reseñas.

Cuadro 2: La proporción de las reseñas de acuerdo a su sentimiento.

Buenos	12500
Malos	12500
Total	25000

Las reseñas también contienen carácteres especiales, así como etiquetas del lenguaje de marcado HTML, por lo que se realizó un análisis de cuáles palabras y carácteres tenían que retirarse del texto, ya que estas no proporcionarían información relevante para el estudio.

3. Preparación de los Datos

La etapa de preprocesamiento es de suma importancia en el Análisis de Sentimientos, y en general en cualquier tarea de Minería de Texto, por lo que es necesario que sea elaborado y ahonde en detalles. A continuación cada una de las etapas de la preparación del texto:

3.1 Limpieza de Datos

Limpiar cada reseña consistió en remover caracteres y etiquetas HTML que no aportaban ningún tipo de información relevante relacionada con las emociones del texto, los elementos removidos fueron: {
, \, / , "}.

Sin embargo, no todos los tipos de etiquetas HTML o caracteres especiales fueron removidos, al igual que se mantuvo las palabras en mayúscula, ya que si presentan un valor importante, dado que destacan palabras de vital importancia semántica a la hora de la clasificación de las reseñas, éstos son: {, <i>, emoticones, signos de exclamación}.

3.2 Tokenización

Tokenizar se refiere a la tarea de dividir cada reseña en sus partes constituyentes, es decir, dividirla en unidades lingüisticas como palabras, signos de puntuación, etc. Estas unidades se denominan *tokens*. No hay una sola forma de tokenizar, el algoritmo apropiado depende de la aplicación [2].

Dado el formato de cada reseña en el dataset y lo que comentamos en el apartado anterior, un token puede ser:

- Palabras.
- Signos de puntuación
- Etiquetas HTML
- Signos de puntuación + Letras, conmunmente conocidos como *emoticones*, siendo relevantes ya que gracias al impacto de las redes sociales, muchas personas complementan sus comentarios agregando formas gráficas de expresión.

El tokenizador usado para este paper se llama *TweetTokenizer* encontrado en la libreria de lenguaje natural NLTK [8].
Ahora, ya tenemos cada una de las reseñas convertidas en un *Bag-of-Words*, ahora necesitamos tomar los tokens más importantes, es decir, eliminar las palabras sin significado relevante, también denominadas *Stop Words*, comúnmente como artículos, pronombres, preposiciones, etc.

3.3 Selección de Características

Última instancia en la preparación del texto. Juega un rol fundamental en la clasificaciíon, ya que esto se refiere a un enfoque que define la manera en que esas características se van a utilizar para clasificar nuevos datos para el tipo específico de clase [3].

Se desea obtener un *Words-Vector*, que se formará con elementos proveniente de la composición de los tokens iniciales generando tokens más complejos, esto se denomina *N-gram*, donde se toma *N* tokens para formar solo un nuevo token, donde $N \geq 1$, siendo el caso base $N = 1$, denominado *Unigram* es el ideal en cuanto a resultados para análisis de sentimientos y es el usado en este paper.

Una vez definido el enfoque que seguirá la interpretación de los tokens, finalmente queda es asignarle un valor numérico a cada token. En este paper hemos experimentado con dos modelos de *Words-Vector*:

- *Frecuencia de Palabras*: dado una reseña, cuenta la cantidad de apariciones de cada token en un corpus.
- *Frecuencia de Palabras – Frecuencia inversa de documento*: dado una reseña cuenta la cantidad de apariciones de cada token dado un corpus, sin embargo le asigna a cada token un peso de acuerdo a la proporción, de dicho token en todo el corpus, a mayor peso, mayor relevancia.

4. Enfoque Propuesto: Random Forest

En esta sección hablaremos acerca de las razones por la que hemos elegido un enfoque con Random Forest, además de cómo hemos ajustado este algoritmo a el caso de estudio.

En Random Forest [4], que fue el primer paper en el que se combinó múltiples árboles de decisión, se explica que este clasificador proporciona dos tipos de aleatoriedad: con respecto a los datos y con respecto a las características, usando el concepto de *Bagging* y *Boostrapping*. A diferencia de usar solo un *Árbol de Decisión*, Random Forest es un clasificador muy robusto al ruido debido a la aleatoriedad que proporciona [3] al ser un *Ensemble Method*, es decir, combina los resultados de múltiples clasificadores para aumentar la precisión de la predicción.

Random Forest trabaja con tres parámetros:

- Números de árboles a construir en el *Decision Forest*.
- Número de características seleccionadas aleatoriamente.
- Profundidad de todos los árboles.

Cada uno de los parámetros tienen su propia importancia e influencia hacia la predicción generada. El *Número de árboles* linealmente aumenta la precisión del modelo. 46 Mientras mayor sea el tamaño de los bosques, mayor es la exactitud, pero la precisión convergerá. En los experimentos realizados, hemos probado con diversos valores de *Número de árboles* para validar esta premisa. La profundidad del árbol también es importante, si se usa un valor muy pequeño, el modelo sufrirá *Under-Fitting*.

En líneas generales, este algoritmo es bastante robusto, preciso y sencillo de manejar y entender su funcionamiento, por lo que hemos elegido este enfoque con respecto a soluciones típicas del área como Naïve Bayes o Máxima Entropía.

5. Experimentos y Resultados

Los resultados obtenidos de los experimentos realizados se encuentran en las siguientes tablas, para esto los datos originales se dividieron en *Entrenamiento* y *Prueba* usando muestreo estratificado con validación cruzada para mantener las clases balanceadas.
Se realizaron en el lenguaje Python con la ayuda de la librería de *Machine Learning*, SciKit-Learn[7]

Cuadro 4: Métricas de evaluación del modelo.

Trees	Accuaracy		Precision		Recall		F1–Score	
10	0.77	0.78	0.77	0.79	0.77	0.78	0.77	0.78
50	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.83
100	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84

En la tabla anterior se puede apreciar los valores obtenidos al evaluar el modelo con el conjunto de *Prueba*, es claro que a medida que aumenta la cantidad de árboles de decisión el *Accuaracy* aumentó, al igual que las otras medidas.

Cuadro 6: Resultados de la clasificación.

Trees	Good–Good		Bad–Bad		Good–Bad		Bad–Good	
10	2216	2268	2578	2629	574	496	909	857
50	2560	2556	2661	2663	464	462	565	569
100	2601	2586	2666	2692	459	539	524	433

En la tabla anterior se puede apreciar los resultados obtenidos de las proporciones de la clasificación de las reseñas en el conjunto de de *Prueba* tomando diferentes números de árboles de decisión así como con los modelos de *Words-Vector*. Con dichos resultados se ve clara la tendencia, el modelo es mejor clasificando las reseñas que son malas que las buenas. Es claro que usando *Random Forest* es un buen enfoque para comenzar y además da buenos resultados, sin embargo no es el mejor, uno de los mejores enfoque en la actualidad para la minería de texto, es el uso de algoritmos de *Deep Learning*. En futuros trabajos se podría profundizar el área y utilizar este enfoque y comparar los resultados obtenidos.

Referencias

[1] Bag of Words Meets Bags of Popcorn. www.kaggle.com/c/word2vec-nlp-tutorial, 2014.

[2] C Potts. Sentiment Symposium Tutorial. sentiment.christopherpotts.net, 2011.

[3] H. Parmar, S. Bhanderi and G. Shah. Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters. *International Conference on Information Science, At Kerala.*, 2014.

[4] L. Breiman. Random Forests. *Machine Learning, vol. 45. Issue 1, pp. 5-32.*, 2001.

[5] C. Aggarwal. Data Mining – The Texbook. *Springer International Publishing*, 2015.

[6] Internet Movie Database. `www.imdb.com`, 2016.

[7] Scikit-Learn, Machine Learning in Python. `scikit-learn.org`, 2016.

[8] Natural Language Toolkit. `www.nltk.org`, 2016.