ISSN 1870-4069

Clasificación de polaridad con un diccionario mediante el algoritmo de Bayes

Yuridiana Alemán, Darnes Vilariño, Josefa Somodevilla

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, México

yuridiana.aleman@gmail.com, darnes@cs.buap.mx, mariajsomodevilla@gmail.com
http://www.cs.buap.mx/

Resumen. En este artículo se presentan una serie de variantes para el algoritmo de Naïve Bayes aplicado a texto. Las variantes se enfocan en la clasificación de polaridad con el uso de un diccionario de palabras que se consideran como positivas o negativas. Entre las propuestas presentadas se utiliza una versión binaria y una ponderada de dicho diccionario. Las pruebas se realizan con dos corpus diferentes, obteniendo mejores resultados con el enfocado a opiniones de foros; además, el uso del diccionario aumenta la exactitud del clasificador y las métricas para la clase negativa.

Palabras clave: Aprendizaje automático, Naïve Bayes, polaridad, diccionario, ponderación.

Polarity Classification with a Dictionary and Bayes Algorithm

Abstract. In this article we present a set of variants for Naïve Bayes algorithm applied to text. The focus are the sentiment polarity using a polarized dictionary. The dictionary specifies if a word is a positive or negative class. Among the presented proposals we use a binary and a weighted version of the dictionary. The experiments are performed with two differents corpus, getting better results with forums focused on opinions; also, dictionary increases the classifier accuracy and metrics for the negative class.

Keywords. Machine Learning, Naïve Bayes, polarity, dictionary, weighting.

1. Introducción

En la actualidad, las redes sociales son la plataforma para poder expresar opiniones, ideas y prácticamente cualquier tipo de información. La cantidad de texto que se genera es tan grande que es imposible analizarlos uno a uno; es por

esto que en la literatura se proponen varias herramientas con diferentes enfoques para poder realizar esta tarea de manera automática. Existen diferentes enfoques para el tratamiento de los textos en socialmedia, una de las más trabajadas es la detección de polaridad (positiva/negativa) de un mensaje, especialmente en la red social de Twitter.

Siguiendo con esta línea de investigación, en el presente trabajo se muestran algunas modificaciones al algoritmo de clasificación de Naïve Bayes para texto analizando dos corpus diferentes y basados en el uso de un diccionario binario.

La estructura del artículo es la siguiente: En la sección 2 se analizan algunos trabajos relacionados con el algoritmo Naïve Bayes y el análisis de polaridad de Twitter, posteriormente, en la sección 3 se explica la metodología a seguir para los experimentos, además, se explica de manera general el algoritmo utilizado. La sección 4 muestra las variantes propuestas para el algoritmo y en la sección 5 se analizan los resultados obtenidos por clasificador y corpus; finalmente, en la sección 6 se presentan las conclusiones obtenidas y el trabajo futuro.

2. Estado del arte

Entre las investigaciones más recientes relacionadas con el análisis de polaridad o con el tratamiento del algoritmo bayesiano que se revisaron para esta investigación se encuentran las siguientes:

En [6] se propone un método para analizar sentimientos en artículos de índole política publicados en socialmedia, aunque también se especifica que es pertinente para otros temas. Como entrenamiento se utilizan artículos de periódicos de Indonesia y se procesan mediante 5 módulos: Lector para determinar la clase del artículo (análisis hecho por humanos), analizador para separar cada palabra de los signos de puntuación, limpiador para eliminar palabras especiales y adverbios, análisis humano para encontrar el pronombre o nombres involucrados en la noticia, además de responder a las preguntas "¿Quién?", "¿A quién?" y "¿Qué?". Las respuestas a estas interrogantes constituirán las palabras clave a tomar en cuenta por el método propuesto. La clasificación se realiza mediante al algoritmo de Bayes, obteniendo 87% de exactitud. Entre los problemas que reportan los autores se encuentran la poca riqueza en vocabulario de los textos, la complejidad del lenguaje utilizado (Indonesio) y la dificultad al localizar con precisión el quién y el qué de cada artículo.

Los autores de [1] categorizan la polaridad en documentos multilingües mediante un nuevo enfoque basado en el meta-aprendizaje genérico. Para esta propuesta realizan la desambiguación del sentido de la palabra, así como algunas funciones basadas en la ampliación del vocabulario. Los autores se enfocan en la combinación de clasificadores y el estudio de los efectos de la desambiguación semántica en el vocabulario de los documentos y por ende, en los atributos de clasificación. Para la desambiguación y obtención de los posibles dominios de las palabras utilizan la herramienta *BabelNet*[1] la cual es un diccionario enciclopédico

[1] `http://babelnet.org/`

multilingüe. Entre los conjuntos de atributos que se analizan se encuentran: TF-IDF bolsa-de-palabras, TF-IDF con n-gramas y un recurso léxico basado en la minería de opiniones; además, se incluye un clasificador independiente para estudiar la contribución de la desambiguación y expansión de conceptos en las etiquetas POS empleadas (adjetivos, sustantivos, verbos y adverbios). Por último, bajo el supuesto de que los conceptos relacionados semánticamente tienen un pariente cercano común, se aprovecha esta posible relación entre los conceptos incluyendo un clasificador basado en la ampliación del vocabulario. Los resultados finales se extraen mediante máquinas de soporte vectorial y C4.5. Los mejores resultados se obtienen con máquinas de soporte vectorial utilizando los conjuntos de atributos que involucran desambiguación semántica.

En [7] se trabaja sobre los enfoques de ponderación de funciones para los clasificadores de Bayes enfocados a texto. La mayoría de los enfoques de ponderación analizados para clasificadores de texto Bayes, la mejora en los resultados implica más tiempo de procesamiento. Por lo tanto, los autores proponen dos enfoques nuevos: Uso de la ganancia de información para la ponderación del texto y ponderar en función de un árbol de decisión. Los resultados experimentales basados en datos de referencia y del mundo real muestran que los enfoques propuestos muestran mayor precisión y mantienen el tiempo de ejecución sin grandes cambios en los modelos finales.

[2] propone un sistema de clasificación automática para el idioma español basada en al análisis del contexto en donde fueron publicados llamado SCOPT. Para la metodología que proponen, se genera un diccionario ponderado con elementos en español con 6 emociones básicas (alegría, sorpresa, repulsión, miedo, enojo y tristeza). El método de clasificación de un tuit es determinado tomando en cuenta la ocurrencia de las palabras del diccionario afectivo así como su ponderación afectiva. La polaridad de un tuit se determina por medio de realizar la combinación lineal de los pesos asignados a cada una de las palabras que aparecen dentro de un tuit y que ocurren dentro del diccionario afectivo. Los resultados alcanzan hasta el 74% de precisión sobre algunos temas.

Finalmente, en [3] se presenta la metodología propuesta para la tarea 4 del SemEval2016, propone el uso de un recurso léxico para el preprocesamiento de datos y entrenar con un modelo de redes neuronales y una representación tipo Doc2Vec. En la metodología se incluyen diccionarios de palabras de argot, contracciones, abreviaturas y emoticones más populares en los medios sociales. Para el proceso de clasificación, se utilizan las características obtenidas sin supervisión en un clasificador SVM obteniendo resultados superiores al *baseline* de la competencia.

3. Metodología

Para la realización de los experimentos se utilizaron dos conjuntos con documentos de diferentes longitudes. El número de documentos del conjunto de entrenamiento y de prueba por corpus se muestran en la Tabla 1.

Tabla 1. Conjuntos utilizados para el análisis.

Corpus	Entrenamiento	Evaluación
Opinion	3,608	1,548
Textos	10,000	1,762

El conjunto denominado *Opinion* contiene principalmente textos cortos obtenidos de foros en donde se emite alguna opinión (positiva o negativa) sobre algún producto o aplicación para celular. Como sus textos y el número de instancias es pequeño, el vocabulario es pobre respecto al corpus de *Textos*, el cual contiene más palabras por documento al ser textos que hablan sobre noticias o eventos recientes, pero siempre manteniendo la opinión positiva o negativa del autor.

Los procedimientos realizados son los siguientes (Figura 1).

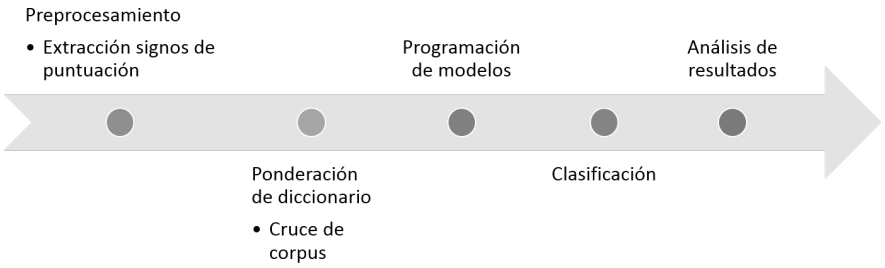

Fig. 1. Proceso de experimentación.

1. Extracción de signos de puntuación, caracteres no imprimibles de las palabras de todos los conjuntos de texto. En este caso no se hizo ningún otro tipo de eliminación ni análisis semántico de las palabras.
2. El diccionario con el que se cuenta contiene una lista de 3,841 palabras y la clase positiva o negativa para cada una de ellas (una palabra no puede ser positiva y negativa al mismo tiempo).
3. Para la programación de modelos, se proponen 4 variantes del clasificador inicial; para estos modelos los atributos son las palabras de los textos.
4. Al final se comparan los resultados obtenidos por corpus y por modelos. La evaluación se realiza con las matrices de confusión y al final un análisis de exactitud.

3.1. Clasificador Naïve Bayes

Todos los experimentos parten del clasificador de Naïve Bayes, el cual es uno de los modelos más simples para clasificación. Se basa en la suposición de que las cantidades de interés son gobernadas por distribuciones de probabilidad y que las

decisiones óptimas pueden estar hechas razonando acerca de estas probabilidades junto con la observación de datos [4]. Se basa en la aplicación del *Teorema de Bayes* para predecir la probabilidad condicional de que una instancia pertenezca a una clase $P(c_i|d_j)$ a partir de la probabilidad de las instancias dada la clase $P(d_j|c_i)$ y la probabilidad *a priori* de la clase en el conjunto de entrenamiento $P(c_i)$ (fórmula 1):

$$P(c_i|d_j) = \frac{P(c_i)P(d_j|c_i)}{P(d_j)}. \tag{1}$$

En el procesamiento de lenguaje natural, esta fórmula se adecua al texto realizando cambios como el uso del logaritmo para evitar probabilidades demasiado pequeñas y la suma de uno para evitar probabilidades de cero (suavizado). Dados estos cambios, en [5] se propone la fórmula 2 :

$$NB(D) = arg_c max log(P(c)) + \sum_{i=1}^{k} log(P(f_i|C)). \tag{2}$$

Todas las variantes presentadas en este artículo se basan en la extracción de $P(f_i|C)$, las cuales se explican en la siguiente sección.

4. Variantes propuestas

Aparte del clasificador usual, se presentan 4 modificaciones más. Las maneras de calcular $P(f_i|C)$ se muestran en la figura 2 y se explican a continuación:

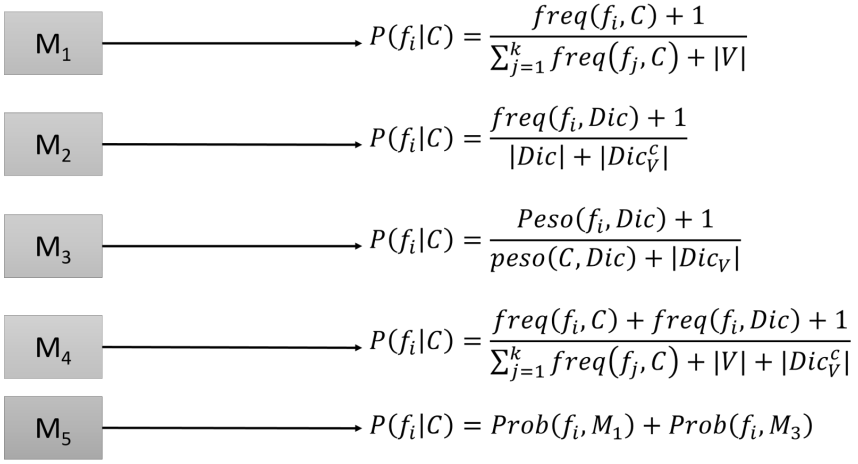

Fig. 2. Variantes propuestas del clasificador.

- M_1**:** Fórmula general para la frecuencia de cada palabra por clase utilizando el suavizado. Cada palabra del conjunto de entrenamiento se convierte en atributo; este modelo fue propuesto en [5].
- M_2**:** Se siguen utilizando palabras como atributos, con la variante de que sólo se utilizarán aquellas que aparezcan en el diccionario. Si una palabra del conjunto de entrenamiento se encuentra en el diccionario y esta tiene la clase del documento analizado, se contabiliza en las frecuencias. Para el caso de las frecuencias 0, se vuelve a utilizar el suavizado y la suma del total de palabras en el denominador de la función. En la fórmula:
  - $frec(f_i, Dic)$: Frecuencia de la palabra en el diccionario, en este caso al ser binario, la frecuencia será 0 si no pertenece a la clase del documento y 1 si pertenece,
  - $|Dic|$: Total de palabras del diccionario que se encuentran en el conjunto de entrenamiento.
  - $|Dic_V^c|$: Total de palabras del diccionario que se encuentran en la clase analizada en el conjunto de entrenamiento.
- M_3**:** Al igual que en M_2, sólo se utilizarán las palabras que aparezcan en el diccionario de acuerdo a la clase analizada, pero en este caso no se trata de un diccionario binario, sino que cada palabra tiene una frecuencia asignada. Esta frecuencia se obtiene mediante el *cruce de corpus*, es decir, se utiliza el corpus *Opinion* para extraer la probabilidad de cada una de las palabras del diccionario (probabilidad simple). Una vez extraída la ponderación de las palabras, ésta se utiliza para calcular el modelo del corpus *Textos*.
- M_4**:** Se combina el diccionario binario con el clasificador usual, es decir, se toma la frecuencia normal de las palabras, pero si la palabra aparece en el diccionario con la misma clase del documento analizado, se le añade uno al cálculo.
- M_5**:** Se combinan el diccionario ponderado con el clasificador usual (M_1 y M_3.

5. Resultados obtenidos

En la Figura 3 se muestran los resultados de la implementación del algoritmo normal con suavizado. Aunque el número de documentos bien clasificados es muy similar, por el número total de instancias los mejores resultados son para el corpus *Opinion* al tener solo un poco más de 100 instancias clasificadas incorrectamente. En ambos corpus se muestran más errores de tipo II (falsos negativos) respecto a la clase positiva.

En la Figura 4 se muestran los resultados para la clasificación utilizando únicamente el diccionario (tanto en su versión binaria como la propuesta de ponderación). Si bien el número de documentos clasificados correctamente baja considerablemente, hay que tomar en cuenta que aún se siguen clasificando bien, más del 50% de las instancias con menos atributos.

Se continúa manteniendo el comportamiento de mayor número de documentos positivos clasificados incorrectamente, además de que en el corpus *Textos*, la

	Opinion		Textos	
	Positivo	Negativo	Positivo	Negativo
Positivo	680	94	610	271
Negativo	34	741	148	734

Fig. 3. Matrices de confusión por corpus para M_1.

	Opinion		Textos	
M2	Positivo	Negativo	Positivo	Negativo
Positivo	512	262	660	221
Negativo	151	624	540	342
M3				
Positivo	518	256	504	377
Negativo	147	628	404	478

Fig. 4. Matrices de confusión por corpus para M_2 y M_3.

clase negativa baja considerablemente el número de documentos bien clasificados en las dos variantes del diccionario. De manera general, el diccionario binario es un poco más eficiente que el ponderado, aunque esto se puede ver seriamente afectado por la calidad del corpus que se utilice para ponderarlo.

En la Figura 5 se muestran los resultados de unir la frecuencia simple de las palabras con el diccionario binario (M_4) y el ponderado (M_5).

	Opinion		Textos	
M4	Positivo	Negativo	Positivo	Negativo
Positivo	686	88	609	272
Negativo	34	741	148	734
M5				
Positivo	688	86	610	271
Negativo	34	741	149	733

Fig. 5. Matrices de confusión por corpus para M_4 y M_5.

Los resultados siguen manteniendo comportamientos similares, pero se observa un ligero incremento en los valores, lo cual hace que se superen los datos inicialmente obtenidos con M_1. A diferencia del uso de los diccionarios solamente, en el corpus *Textos* mejoró considerablemente la clasificación de los textos negativos; además, se incrementaron los valores de la diagonal principal. Para visualizar de manera general las modificaciones propuestas, en la Figura 6 se muestra la exactitud obtenida por modelo propuesto y corpus.

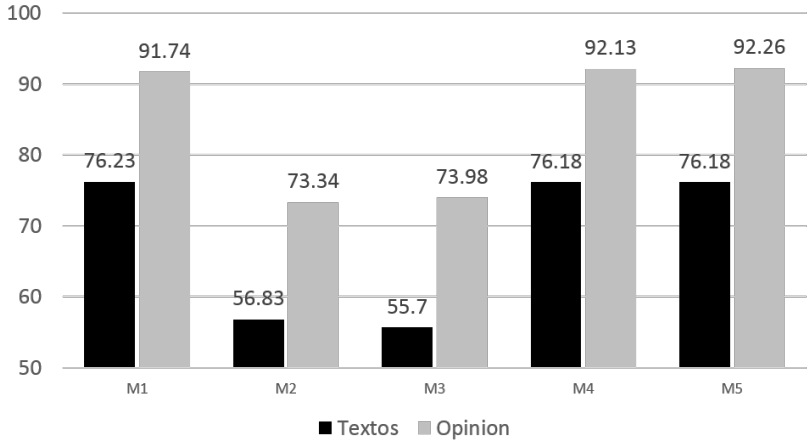

Fig. 6. Variantes propuestas del clasificador.

Si bien el algoritmo tradicional supera por ejemplo a las dos implementaciones del diccionario, en el caso de la fusión de los modelos hay un incremento que aunque muy pequeño, se vuelve importante al manejar grandes cantidades de documentos. Para el caso del corpus *Textos* no se llegó a una mejora, pero con el corpus *Opinion* se logró mejorar la calidad del clasificador

6. Conclusiones y trabajo futuro

En el presente artículo se presentaron varias propuestas para incluir un diccionario en el clasificador de Naïve Bayes. Si bien sólo en un corpus se obtuvo un ligero incremento se pueden llegar a las siguientes conclusiones:

- Tanto en los modelos de M_4 y M_5 se encuentran mejoras en la clasificación de los documentos positivos, especialmente en el corpus *Opinion*.
- El diccionario ponderado mejora poco o nada la exactitud del clasificador, por lo que se debe analizar otras formas de ponderación que incrementen esta métrica.

– Aunque en este caso se realizaron experimentos para clases de tipo positivo-negativo, la metodología es aplicable a cualquier clasificación binaria, siempre y cuando se cuente con la lista de palabras representativas de cada clase.
– Este trabajo se puede considerar como *baseline* para la experimentación futura con otros corpus e incluso verificar la fiabilidad del diccionario por medio de técnicas estadísticas.

Referencias

1. Franco-Salvador, M., Cruz, F.L., Troyano, J.A., Rosso, P.: Cross-domain polarity classification using a knowledge-enhanced meta-classifier. Knowledge-Based Systems 86, 46 – 56 (2015), http://www.sciencedirect.com/science/article/pii/S0950705115002063
2. Gálvez-Pérez, J.R., Gómez-Torrero, B.E., Ramírez-Chávez, R.I., Sánchez-Sandoval, K.M., Castellanos-Cerda, V., García-Madrid, R., Jiménez-Salazar, H., Villatoro-Tello, E.: Sistema automático para la clasificación de la opinión pública generada en twitter. Research in Computing Science 95, 23–36 (2015), http://rcs.cic.ipn.mx/2015_95/Sistema%20automatico%20para%20la%20clasificacion%20de%20la%20opinion%20publica%20generada%20en%20Twitter.pdf
3. Gómez-Adorno, H., Vilariño, D., Sidorov, G., Avendaño, D.P.: Cicbuapnlp at semeval-2016 task 4-a: Discovering twitter polarity using enhanced embeddings. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016. pp. 145–148 (2016), http://aclweb.org/anthology/S/S16/S16-1021.pdf
4. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. pp. 338–345. UAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995)
5. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
6. Soelistio, Y.E., Surendra, M.R.S.: Simple text mining for sentiment analysis of political figure using naive bayes classifier method. CoRR abs/1508.05163 (2015), http://dblp.uni-trier.de/db/journals/corr/corr1508.html#SoelistioS15
7. Zhang, L., Jiang, L., Li, C., Kong, G.: Two feature weighting approaches for naive bayes text classifiers. Knowledge-Based Systems 100, 137 – 144 (2016), http://www.sciencedirect.com/science/article/pii/S0950705116001039