Aplicación del modelo Léxico-Sintáctico para detectar la polaridad de opiniones sobre profesores

Darnes Vilariño
Benemérita Universidad Autónoma de Puebla: BUAP
yuvisrosas@gmail.com

Claudia Zepeda
Benemérita Universidad Autónoma de Puebla: BUAP
czepedac@gmail.com

Yuvila M. Sanzón
Benemérita Universidad Autónoma de Puebla: BUAP
dvilarinoayala@gmail.com

José L. Carballido
Benemérita Universidad Autónoma de Puebla: BUAP
jlcarballido7]@gmail.com

Carolina Medina
Universidad Autónoma Metropolitana, Unidad Iztapalapa: UAM-I
rcarolinamedina@gmail.com

Georgina Flores
Instituto Tecnológico de Puebla: ITP
kremhilda@gmail.com

Resumen

En el presente trabajo se muestran los resultados obtenidos de la aplicación del modelo léxico sintáctico a las opiniones dadas por estudiantes de la Facultad de Ciencias de la

Computación de la Benemérita Universidad Autónoma de Puebla, acerca de los profesores que impartieron cursos en verano del 2015. Se confeccionó un corpus categorizado con las opiniones obtenidas de una encuesta aplicada. El corpus obtenido permitió la confección de un modelo de clasificación que permite detectar la polaridad de opinión (positiva, negativa o neutra). Los resultados obtenidos desarrollando el modelo con el 80% de las opiniones y probando con el 20% ofrecieron una precisión del 65%.

Palabra(s) Clave(s): análisis de sentimientos, minería de opinión, modelo léxico sintáctico, proceso enseñanza-aprendizaje.

1. Introducción

Descubrir el sentimiento que expresa un individuo ante un determinado servicio que recibe, se ha vuelto muy importante, ya que es necesario desarrollar mecanismos continuos para la mejora de cualquier proceso, a esto se le conoce como minería de opinión [11]. Parte importante de la minería de opinión es la detección de la polaridad del texto que se está escribiendo [12].

En la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla (FCC-BUAP), existe un índice de deserción cercano al 25% en los últimos años [10,11], esto nos ha motivado a investigar las causas que están provocando esta problemática. La investigación se llevó a cabo mediante una encuesta y se dividió en dos partes. En la primera parte se le pidió a cada estudiante su opinión con respecto a los profesores que les impartieron clase en el verano del 2015, sin tomar en cuenta la materia ni el nombre del profesor. En la segunda parte se decidió analizar si el problema está asociado directamente con la disciplina de Matemáticas, por lo que esta parte corresponde a un conjunto de 24 preguntas de opción múltiple relacionadas con el proceso de enseñanza-aprendizaje de las matemáticas y los antecedentes de los estudiantes en los cursos de matemáticas. El presente trabajo reporta los resultados obtenidos solo de la primera parte de la encuesta.

Con el fin de mejorar el proceso enseñanza-aprendizaje en la FCC-BUAP, este trabajo da un paso inicial hacia el desarrollo de una herramienta que teniendo como entrada una opinión, esta opinión pueda ser clasificada como positiva, negativa o neutra. Por lo tanto, la intención de este trabajo es además conocer si el modelo léxico-sintáctico propuesto e implementado en [7] podría formar parte de la mencionada herramienta. El modelo léxico-sintáctico es capaz de descubrir la polaridad de un mensaje indicando si el mensaje es positivo, negativo o neutro. En particular, en [7] se exponen los resultados obtenidos del modelo léxico-sintáctico con los datos ofrecidos en el marco de la competencia SemEval 2014, en particular para mensajes en idioma Inglés de Tweeter.

En este artículo se describe la adaptación que se tuvo que hacer al sistema que implementa el modelo léxico-sintáctico, para procesar las opiniones de nuestra encuesta que dio cada uno de los estudiantes sobre sus profesores; los resultados que se obtienen al ser utilizado el mencionado sistema; y finalmente nuestras conclusiones. Cabe destacar que otra de las aportaciones de este trabajo es la confección y categorización de un corpus a partir de las opiniones dadas por los alumnos de la FCC-BUAP. Precisamente este corpus sirvió para probar el modelo léxico-sintáctico propuesto e implementado en [7].

En la sección 2 se presentan los preliminares de este trabajo sobre la encuesta aplicada y una descripción general del modelo léxico-sintáctico. En la sección 3 se describe la metodología seguida para el desarrollo de este trabajo. La sección 4 muestra los resultados obtenidos y finalmente en la sección 5 las conclusiones a las que llegamos.

2. Preliminares

Como se ha dicho en la sección de Introducción, en la FCC-BUAP, existe un índice de deserción cercano al 25% en los últimos años, lo cual se puede verificar en la Tabla 1[1], en esta tabla CCO, ICC, ITI corresponden a los nombres de los tres programas

[1] Información proporcionada por la Secretaría Académica de la FCC-BUAP el día 13 de agosto del 2015.

educativos que se imparten en la FCC: Licenciatura en Ciencias de la Computación, Ingeniería en Ciencias de la Computación, e Ingeniería en Tecnologías de la Información.

Generación	Periodo	Clave	% Deserción
2009	Otoño 2009	CCO	35
2009	Otoño 2009	ICC	17
2010	Otoño 2010	CCO	27
2010	Otoño 2010	ICC	22
2011	Otoño 2011	CCO	25
2011	Otoño 2011	ICC	27
2012	Otoño 2012	CCO	27
2012	Otoño 2012	ICC	22
2012	Otoño 2012	ITI	24

Tabla 1. Porcentaje de deserción de la FCC del 2009 al 2012.

Lo anterior nos motiva a investigar las causas que están provocando esta problemática. Para estudiar esta problemática, se desarrolló una encuesta que se dividió en dos partes[2]. En la primera parte se le pidió a cada estudiante su opinión con respecto a los profesores que les impartieron clase en el verano del 2015, sin tomar en cuenta la materia ni el nombre del profesor. En la segunda parte se decidió analizar si el problema está asociado directamente con la disciplina de Matemáticas. El presente trabajo ocupa los resultados obtenidos solo de la primera parte de la encuesta. Por tanto, esta sección describe la encuesta aplicada y presenta una breve descripción del Modelo Léxico-

[2] La encuesta completa puede verse en https://goo.gl/zy4Zo1

Sintáctico propuesto en [7] y utilizado en este trabajo para detectar la polaridad de las opiniones emitidas por los estudiantes.

2.1 Encuesta aplicada

La primera parte de la encuesta que se aplicó a 575 estudiantes de una población de alrededor de 2000 en verano 2015 y es la siguiente:

Opina sobre tus profesores en verano 2015

Instrucciones: Responde a la siguiente pregunta con un párrafo que describa tu opinión por cada curso que estés tomando en Verano del 2015. Se muestran tres espacios, uno por cada curso que estés tomando.

1. Qué opinión tienes del profesor que imparte uno de los cursos de verano 2015?

_____

2. Qué opinión tienes del profesor que imparte tu segundo curso de verano 2015?

_____

3. Qué opinión tienes del profesor que imparte tu tercer curso de verano 2015?

_____

2.2 Modelo léxico sintáctico

Se han desarrollado distintos trabajos en el área de análisis de sentimientos [1,2,3,4,5,6,8,9], la mayoría de estos se han centrado en el idioma inglés, dado a que existe una gran cantidad de herramientas de procesamiento del lenguaje natural

disponibles y existen conjuntos de datos que pueden ser usados para el entrenamiento y creación de los modelos de clasificación.

En el trabajo desarrollado en [7] se propone un modelo léxico-sintáctico que consta de tres fases, para descubrir la polaridad de los mensajes (positivo, negativo o neutro), este modelo fue desarrollado en Python con ayuda de las herramientas Network X y CLIPS Pattern. La primera fase normaliza los textos utilizando diccionarios léxicos, la segunda fase desarrolla el modelo de clasificación y la tercera fase es la etapa de prueba de dicho modelo. Véase Fig. 1. En la fase de normalización [7] se realiza el pre-procesamiento de los datos de entrenamiento y de los datos de prueba; al final de esta fase se obtienen los archivos de entrenamiento y prueba que son utilizados en la fase de entrenamiento. En la fase de entrenamiento se utilizan los clasificadores Naïve Bayes y Máquina de Soporte Vectorial (SVM), proporcionados por la herramienta CLIPS Pattern. El modelo de clasificación es desarrollado con cada uno de los textos contenidos en el archivo de entrenamiento normalizado, este modelo se utiliza para clasificar los datos en la fase de prueba.

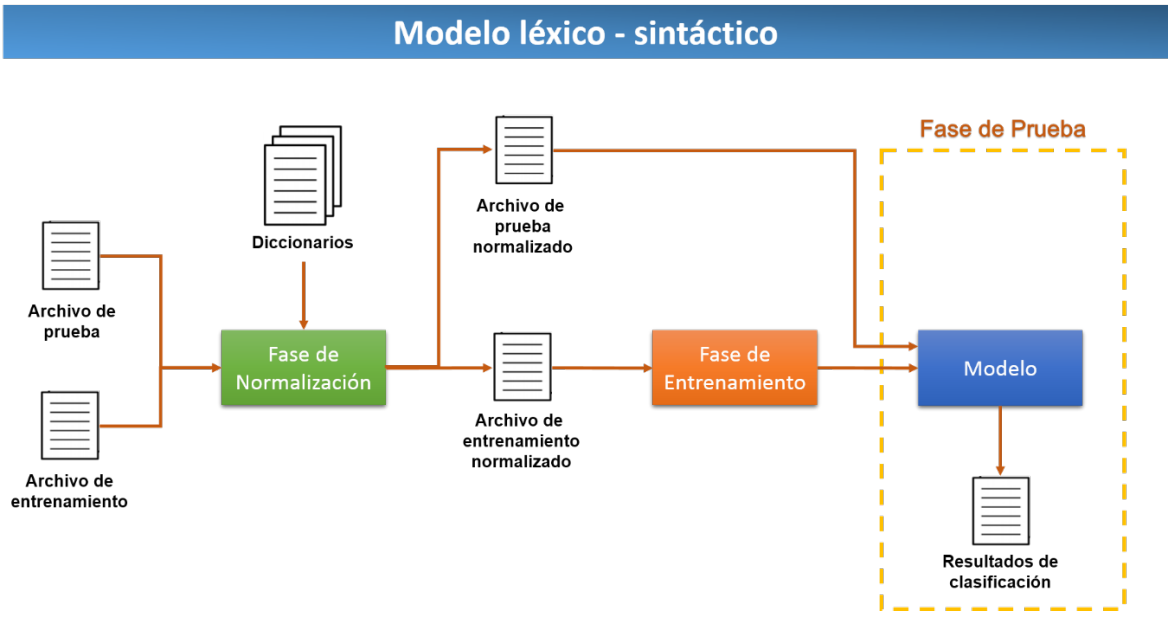

Fig. 1. Arquitectura del modelo léxico – sintáctico.

3. Metodología

Cada opinión dada por los estudiantes fue categorizada por tres expertos, lo que permitió, conformar el corpus de entrenamiento.

Para aplicar el modelo desarrollado en [7] fue necesario realizar adaptaciones sobre todo en la fase de normalización. En la mencionada fase de normalización, se tuvo que modificar el procedimiento dado a que los comentarios están en idioma Español. Iniciando por reemplazar ciertos caracteres como: ñ – ni, á – a, é – e, í – i, ó – o, ú – u. Se eliminaron las stopwords (palabras sin significado), modificando el código para eliminar éstas palabras en idioma español. Para la fase de entrenamiento, los comentarios que se encontraban en un archivo .xlsx se pasaron a un archivo .txt el cual tiene el formato clase (sentimiento: positivo, neutral o negativo) separados por un tabulador seguido del comentario.

Del corpus categorizado, se seleccionó el 80% de los comentarios para el desarrollo del modelo de clasificación y el 20% restante de las opiniones permitió probar la precisión obtenida por el modelo desarrollado. Este 20% de las opiniones conforma el corpus de prueba.

En la Tabla 2, se muestra la composición del corpus de entrenamiento.

Composición del Corpus		
Positivos	**Neutros**	**Negativos**
423	62	65

Tabla 2. Composición del corpus.

Como puede apreciarse, el corpus obtenido está totalmente desbalanceado, es sabido que los modelos de clasificación son muy sensibles a esta situación, es por ello que se toma el 80% de los comentarios basándose en la clase con menor número de instancias,

en este caso fue la clase neutral. En la Tabla 3 se muestra el número de opiniones consideradas en cada clase.

Corpus para el entrenamiento		
Positivos	**Neutros**	**Negativos**
62	62	62

Tabla 3. Composición del corpus de entrenamiento.

En fase de prueba, en cuanto al corpus de prueba se tomaron los comentarios restantes. Ver Tabla 4. Como se sabe cuál es la categoría de cada opinión, con este corpus se puede medir la precisión lograda con el modelo de clasificación.

Corpus de prueba		
Positivos	**Neutros**	**Negativos**
373	12	19

Tabla 4. Composición del corpus de prueba.

En la próxima sección se discuten los resultados obtenidos del modelo Léxico-Sintáctico utilizando los dos clasificadores.

4. Resultados del modelo

A pesar, de que el número de muestras en los datos de entrenamiento es pequeño, el modelo desarrollado utilizando el clasificador Naïve Bayes obtiene para los datos de

prueba un 63.36% de precisión. Este resultado puede ser mejorado si se enriquece el corpus de entrenamiento con nuevas muestras de cada una de las clases. Con esta herramienta propuesta, esto puede lograrse, ya que dada una opinión cualquiera, de manera automática se puede obtener la polaridad de la misma.

Los resultados obtenidos con los datos de prueba utilizando el modelo de clasificación desarrollado con el clasificador Naïve Bayes se muestran en la tabla 5. Puede observarse, que de las 373 opiniones con polaridad positiva, el modelo solamente pudo detectar 248, le dio la categoría de neutro y negativo a opiniones que realmente eran positivas, lo que nos hace pensar que los alumnos usan palabras similares en mensajes de polaridades distintas.

Positivos	Neutros	Negativos	Total de comentarios	Total de aciertos	Porcentaje de precisión
248	75	81	404	256	63.36%

Tabla 5. Resultados Modelo Léxico – Sintáctico con Clasificador Naïve Bayes.

Los resultados obtenidos con los datos de prueba utilizando el modelo de clasificación desarrollado con el clasificador SVM se muestran en la tabla 6. Se realizaron diferentes experimentos utilizando varios kernel, puede apreciarse que solamente el kernel lineal ofrece resultados aceptables, ya que los kernel polinomial de grado 2, polinomial de grado 3 y Radial clasifican la mayoría de las opiniones de manera negativa. El clasificador SVM, utiliza mayormente la frecuencia de las palabras, lo que corrobora nuevamente que los estudiantes utilizan las mismas palabras y su misma frecuencia para expresar opiniones de polaridad diferente.

Kernel	Positivos	Neutros	Negativos	Total de comentarios	Total de aciertos	Porcentaje de Precisión
Lineal	242	69	93	404	253	62.62%
Polinomial grado 2	0	0	404	404	19	4.70%
Polinomial grado 3	0	0	404	404	19	4.70%
Radial	7	28	369	404	26	6.43%

Tabla 6. Resultados Modelo Léxico – Sintáctico con SVM.

5. Conclusiones

Por los resultados obtenidos, se aprecia que el modelo desarrollado en [7] si puede ser aplicado para conocer la polaridad de la opinión de un estudiante con respecto a sus profesores.

Esta herramienta de manera automática, nos puede servir para clasificar nuevas opiniones y de esta forma enriquecer el corpus de entrenamiento y obtener posteriormente un modelo de clasificación más exacto.

Esto es un primer acercamiento para tener una herramienta automática que realice minería de opinión y que a futuro pudiera aplicarse en los cursos y poder detectar aquellos elementos del proceso de enseñanza-aprendizaje que están provocando que el índice de deserción en la FCC sea alrededor de la cuarta parte de su población.

Se desarrolló un modelo de clasificación utilizando Naïve Bayes y SVM, se obtuvo que el modelo obtenido aplicando SVM dio pésimos resultados, ya que la evidencia para desarrollar el mismo en esta primera prueba fue muy escasa. El clasificador Naive-Bayes logra desarrollar un modelo con mayor precisión.

Agradecimientos

Agradecemos a los alumnos Lady Yedidia Mendez Trejo, Ricardo Alejandre Trigo, Alberto Esteban Reyes Peralta, Jaime David Cardoso Juárez, y David Fragoso Porras por apoyarnos en la aplicación de la encuesta. También agradecemos al programa proyectos VIEP 2015 de la Benemérita Universidad Autónoma de Puebla por el apoyo para la realización de este trabajo.

Bibliografía

[1] C. Levallois. Sentiment Analysis for Tweets based on Lexicons an Heuristics. http://www.cs.york.ac.uk/semeval-2013/accepted/27_Paper.pdf (2013). Accedido el 12 de mayo de 2014.

[2] V. Hangya, G. Berend, R. Farkas. Sentiment Detection on Twitter Messages. http://www.cs.york.ac.uk/semeval-2013/accepted/102_Paper.pdf (2013). Accedido el 12 de mayo de 2014.

[3] Y. Wilks, M. Stevenson. The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation. Journal of Natural Language Engineering, Vol. 4, No. 2, pp. 135-143 (1998).

[4] C. Whitelaw, N. Garg. S. Argamon. Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM ’05, pp. 625-631 (2005).

[5] T. Nasukawa, J. Yi. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. Proceedings of the 2nd international conference on Knowledge capture, K-CAP ‘03, pp. 70-77 (2003).

[6] L. Freeman. Centrality in social networks: Conceptual clarification. Social Networks. Journal: Social Networks – SOC NETWORKS, Vol. 1, No. 3, pp. 215-239 (1979).

[7] Y. Sanzón, D. Vilariño.; C. Zepeda.; Pinto, D.; Tovar, M.: Modelos para Detectar la Polaridad de los Mensajes en Redes Sociales. Por publicarse en Journal of Research in Computing Science 2015.

[8] J. Fernández. Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog. Procesamiento del Lenguaje Natural, [S.l.], v. 47, p. 179-187, sep. 2011. ISSN 1989-7553. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/963. Accedido el 6 de agosto del 2015

[9] J. De Albornoz. Un modelo lingüístico-semántico basado en emociones para la clasificación de textos según su polaridad e intensidad. Tesis Doctoral. Universidad Complutense de Madrid. Madrid, septiembre de 2011. http://nlp.uned.es/~jcalbornoz/papers/PhD_Thesis_2011.pdf. Accedido el 6 de agosto del 2015

[10] Deserción estudiantil. Milenio diario. http://www.milenio.com/puebla/Reporta-BUAP-desercion-estudiantil_0_146385703.html. Accedido el 6 de agosto del 2015

[11] J. Zambrano. Implementa BUAP plan para prevenir deserción escolar. E-consulta.com. Publicado el Martes, Agosto 4, 2015. http://e-consulta.com/nota/2015-08-04/universidades/implementa-buap-plan-para-prevenir-desercion-escolar. Accedido el 6 de agosto del 2015.