Detección de subjetividad en noticias en línea publicadas en español utilizando clasificadores probabilísticos

Noé Alejandro Castro-Sánchez[1], Sadher Abelardo Vázquez-Cámara[1] y Grigori Sidorov[2]

[1] Centro Nacional de Investigación y Desarrollo Tecnológico, México

[2] Instituto Politécnico Nacional, Centro de Investigación en Computación, México

{ncastro, sadhervazquez}@cenidet.edu.mx, sidorov@cic.ipn.mx

Resumen. Los textos periodísticos pueden clasificarse dentro del llamado género informativo si su contenido se orienta a la objetividad (descripción de los hechos ocurridos), o en el género de opinión, si incluye elementos subjetivos (como el punto de vista o ideología del autor de la nota). Uno de los problemas que se presenta en la redacción de noticias es que en las notas de tipo informativo se llegan a incorporar elementos subjetivos sin previa advertencia al lector. En este artículo se presenta un método para la detección automática de subjetividad en oraciones de noticias escritas en español. Se generó un corpus a partir de noticias publicadas en internet, las cuales contienen 8,108 oraciones que se clasificaron manualmente como objetivas (3,648) y subjetivas (4,460). Los mejores resultados obtenidos a partir de experimentos con diversos clasificadores automáticos arrojan un 76.3% de precisión, utilizando el clasificador *Bayes Net*.

Palabras clave: detección automática de subjetividad, detección de opinión, corpus de noticias, Naive Bayes, Bayes Net, Weka.

1. Introducción

La cantidad de usuarios de internet en México es de aproximadamente 51.2 millones de personas (más de la mitad de la población), las cuales en su mayoría oscilan en las edades entre 19 y 44 años. Entre los principales usos de internet en México se encuentra la recepción y envío de correos electrónicos, la búsqueda de información y el acceso a redes sociales [1].

Actualmente, México se encuentra en un cambio en cuanto a preferencias de uso de medios de comunicación, en donde dominaba el uso de la televisión y el uso de Internet se encuentra en constante aumento. El uso de internet es la principal fuente de noticias e información, seguida de periódicos, televisión y radio [2].

La subjetividad en cuanto al lenguaje natural hace referencia a aspectos del lenguaje utilizados para expresar opiniones y evaluaciones [3]. El lenguaje subjetivo es un tipo

de lenguaje utilizado para expresar estados privados, los cuales son términos que cubren opiniones, evaluaciones, emociones y especulaciones [4].

De manera general, los géneros periodísticos pueden ser clasificados como géneros informativos, en donde la información de algún hecho o dato se presenta tal y como ha ocurrido, dominando el uso impersonal y objetivo del lenguaje, y los géneros de opinión, en donde el escritor expresa su punto de vista acerca de un hecho o dato, dominando la subjetividad [5].

El papel que juegan los medios de comunicación en la política de México es reciente, y estos desempeñan dos papeles esenciales: Funcionan como diseminadores de información, lo cual es pieza clave en toda democracia y pueden movilizar la opinión pública, así como generar diferentes formas de actividad política [6].

Uno de los problemas que se presentan en el periodismo es que existen noticias que se publican dentro del género informativo, aunque en realidad incluyen opiniones de los escritores de la nota.

El objetivo de este trabajo, es desarrollar un método de detección de subjetividad en oraciones de noticias. Este método consiste en realizar una clasificación automática utilizando clasificadores probabilísticos y un corpus de noticias publicadas en español de México.

En este artículo, se presentan las pruebas y resultados obtenidos al utilizar los clasificadores probabilísticos con el corpus etiquetado de manera manual, realizando diversos experimentos, con el fin de determinar cuál es el mejor clasificador y el mejor conjunto de características.

El artículo se encuentra organizado de la siguiente manera: En la sección 2, se presentan los trabajos relacionados, en la sección 3 se presenta nuestra metodología de solución, describiendo el proceso realizado, presentando los resultados obtenidos en la sección 4. Para finalizar, se presentan las conclusiones y el trabajo a futuro en la sección 5.

2. Trabajos relacionados

Los trabajos que implementan diversas técnicas para la detección de subjetividad son tratados a continuación.

En el trabajo [7], se presenta un marco de trabajo utilizado para identificar declaraciones subjetivas en títulos de noticias, utilizando los sentidos de palabras para identificar el significado (objetividad) y sentido (subjetividad) de cada oración, además de determinar la emoción expresada. Los resultados de precisión de este trabajo muestran un 73% de precisión, aunque al combinarse con características extras, puede llegar a obtenerse un 99%.

En el trabajo [8], se propone realizar minería de opinión para detectar opiniones en columnas de noticias de género político publicadas en idioma tailandés. El trabajo se compone de 3 partes: Colección de datos (limpieza y almacenamiento de información), anotación (etiquetado manual por humanos) y clasificación (minería de datos con *Weka*). Los mejores resultados muestran una precisión del 80.7% al utilizar el clasificador *Naive Bayes*.

Otro trabajo es *OpinionFinder* [9], un sistema de análisis de subjetividad que identifica de manera automática cuando existen opiniones, sentimientos, especulaciones y otros estados privados en el texto. Opera en 2 partes: En la primera parte realiza un procesamiento del documento y en la segunda parte, se realiza el análisis de subjetividad, que consta de 4 componentes: clasificación de sentencias subjetivas, eventos del discurso y clasificación de expresión subjetiva directa, identificación del origen de opinión y por último, clasificación de la expresión del sentimiento.

En el trabajo [10], se presenta *SubjLDA*, un modelo jerárquico bayesiano basado en Asignación *Dirichlet* latente (*Latent Dirichlet Allocation*, *LDA*), para la detección de subjetividad a nivel de oraciones, el cual automáticamente identifica si una oración dada expresa opiniones o si expresa hechos. El proceso generativo involucra tres etiquetas de subjetividad para las sentencias, una etiqueta de sentimientos para cada palabra en la oración y las palabras en las oraciones. El algoritmo de *subjLDA* es un modelo bayesiano de cuatro capas y la clasificación de la subjetividad en la sentencia es determinada directamente desde la etiqueta de subjetividad de la sentencia. En los resultados obtenidos, se demostró que *subjLDA* obtuvo un porcentaje de precisión del 71.6% al analizar sentencias objetivas y un 71% al analizar sentencias subjetivas, dando un resultado final de precisión de precisión de 71.2%.

3. Método propuesto

Nuestro trabajo propone un método de detección de subjetividad de noticias utilizando un clasificador automático y tomando características a nivel de oración y n-gramas. La figura 1 muestra la arquitectura del método:

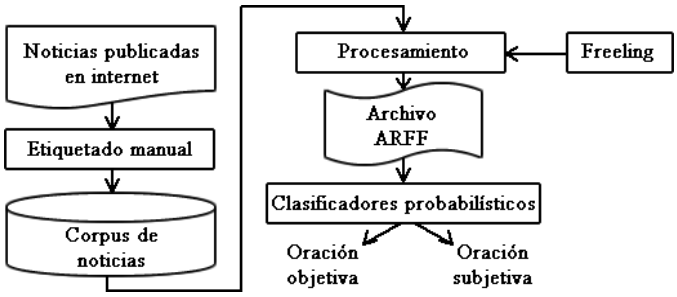

Fig. 1. Arquitectura del método de detección de subjetividad.

Primero, se obtienen noticias publicadas en internet, las cuales son etiquetadas de manera manual a nivel de oración por etiquetadores humanos, con el fin de generar el corpus de noticias. Después, se realiza el procesamiento de texto (generación de n-gramas, lematización) para crear el archivo ARFF que servirá de entrenamiento a los clasificadores probabilísticos. Finalmente, se realizan experimentos con diferentes clasificadores para determinar con cuál de estos se obtienen los mejores resultados de detección de objetividad y subjetividad en las oraciones.

3.1. Creación del corpus de noticias

Se creó un corpus de noticias escritas en español obtenidas de diversos sitios periodísticos mexicanos, con el fin de utilizarlas como datos de entrenamiento para los clasificadores probabilísticos. El corpus cuenta con un total de 1972 noticias, las cuales fueron divididas en 2 secciones: Noticias informativas, con un total de 1834 y noticias de opinión con un total de 138. Las notas informativas se conforman por noticias relacionadas a temas de interés sociopolítico (por ejemplo, política, seguridad y economía), descartando notas asociadas a entretenimiento (espectáculos, deportes, etc). Las oraciones de las noticias informativas fueron etiquetadas de manera manual a nivel de oración. Los etiquetadores recibieron una inducción acerca de objetividad y subjetividad, la cual se basó en un manual de etiquetado, desarrollado con base en el trabajo de [11] y [4]. Por otro lado, todas las oraciones de las noticias de opinión fueron etiquetadas como subjetivas. Algunas de las características que se tomaron en cuenta para determinar si una oración es subjetiva, son la modalidad oracional, léxico valorativo, signos de puntuación, entre otros.

Las oraciones de las noticias ya clasificadas arrojaron un total de 3648 oraciones objetivas y 4460 oraciones subjetivas.

3.2. Preprocesamiento de datos para experimentos

Para poder realizar los experimentos con clasificadores probabilísticos, se etiquetó el corpus de manera manual a nivel de oración, obteniendo un total de 8108 oraciones, de las cuales 3648 fueron clasificadas como objetivas y 4460 como subjetivas. Para realizar las pruebas, se generó un único archivo de extensión ARFF con el contenido de todas las notas etiquetadas, con el fin de utilizar los clasificadores que provee el software de minería de datos *Weka* [12]. El formato del archivo ARFF es el siguiente: el atributo "clasificación" indica si la oración fue etiquetada como objetiva (O) o subjetiva (S), mientras que el atributo "texto" contiene la oración o el n-grama, dependiendo del caso.

```
@relation OBJ_SUB
@attribute clasificacion {O,S}
@attribute texto String
@data
O, 'El caso inició el sábado 25 de octubre, cuando la joven
solicitó vía telefónica el apoyo de las autoridades'
S, 'Dicen que el funcionario público es el verdadero culpable
del incidente'
```

Durante el preprocesamiento de los datos, se utilizó el filtro *StringToWordVector*, el cual se encarga de convertir atributos de tipo cadena (*String*) en un conjunto de atributos, los cuales representan la ocurrencia de palabras dentro del texto contenido en la cadena. Los clasificadores utilizados en este trabajo de investigación fueron *Naive Bayes* y *Bayes Net*, pues según consta en la bibliografía son los que mejores resultados arrojan [13] [14]. Se realizaron diversos experimentos para identificar las mejores

características para los clasificadores, los cuales pueden dividirse de la siguiente manera según las diferentes combinaciones derivadas del tratamiento del texto:

1. Utilización de oraciones.
   (a) Inclusión de *stopwords* (palabras auxiliares) con texto sin lematizar,
   (b) Inclusión de *stopwords* con texto lematizado,
   (c) Eliminación de *stopwords* con texto sin lematizar,
   (d) Eliminación de *stopwords* con texto lematizado.

2. Segmentación de las oraciones en bigramas, trigramas y 4-gramas, a los cuales se les asigna automáticamente la etiqueta de la oración de donde son extraídos. El procesamiento realizado en el punto anterior también se aplicó para cada uno de los siguientes criterios:
   (a) Eliminación de n-gramas repetidos, dejando un único n-grama,
   (b) Eliminación de n-gramas etiquetados como subjetivos, que estuvieran etiquetados también como objetivos,
   (c) Eliminación de n-gramas etiquetados como objetivos, que estuvieran etiquetados también como subjetivos,
   (d) Eliminación de n-gramas etiquetados tanto como objetivos y subjetivos.

4. Resultados de experimentos

A continuación, se muestran los resultados obtenidos utilizando los criterios indicados a nivel oración:

En la tabla 1, podemos observar los resultados obtenidos al utilizar oraciones con el clasificador *Naive Bayes*. La mayor precisión al determinar objetividad fue 65%, en cuanto a subjetividad, la mayor precisión fue 74%. Las características que se consideraron fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 1. Resultados de experimentos utilizando oraciones con el clasificador *Naive Bayes*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
CSW, TNL	**0.65**	0.73	0.68	0.70	0.66	0.71
CSW, TL	0.60	0.71	0.68	0.62	0.64	0.66
SSW, TNL	0.63	**0.74**	0.71	0.67	0.67	0.70
SSW, TL	0.60	0.73	0.72	0.60	0.66	0.66

En la tabla 2, se observan los resultados obtenidos al utilizar oraciones con el clasificador *Bayes-Net*. La mayor precisión al determinar fue 63%, y en subjetividad fue 70%. Las características que se tomaron fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 2. Resultados de experimentos utilizando oraciones con el clasificador *Bayes Net*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
CSW, TNL	0.60	0.66	0.57	0.69	0.58	0.68
CSW, TL	**0.63**	0.59	0.28	0.86	0.39	0.70
SSW, TNL	0.61	**0.70**	0.66	0.65	0.63	0.67
SSW, TL	0.59	0.59	0.28	0.84	0.38	0.69

Después, se realizaron experimentos, eliminando n-gramas repetidos, dejando solamente un n-grama en el archivo ARFF.

En la tabla 3, se puede observar los resultados obtenidos al utilizar el clasificador *Naive Bayes.* La mayor precisión al determinar objetividad fue de 63% y al determinar subjetividad la mayor precisión fue de 67%. En este caso, las características fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 3. Resultados de experimentos utilizando n-gramas, eliminando n-gramas repetidos con el clasificador *NaiveBayes*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.52	0.53	0.36	0.68	0.42	0.60
Bigramas, CSW, TL	0.62	0.52	0.11	0.93	0.18	0.67
Bigramas, SSW, TNL	0.53	0.52	0.29	0.74	0.38	0.61
Bigramas, SSW, TL	**0.63**	0.57	0.42	0.76	0.50	0.65
Trigramas, CSW, TNL	0.54	0.55	0.52	0.56	0.53	0.55
Trigramas, CSW, TL	0.56	**0.67**	0.79	0.40	0.66	0.50
Trigramas, SSW, TNL	0.54	0.54	0.48	0.60	0.51	0.57
Trigramas, SSW, TL	0.56	0.63	0.73	0.45	0.63	0.52
4-gramas, CSW, TNL	0.56	0.56	0.58	0.54	0.57	0.55
4-gramas, CSW, TL	0.55	0.66	0.80	0.37	0.66	0.47
4-gramas, SSW, TNL	0.56	0.57	0.60	0.52	0.58	0.54
4-gramas, SSW, TL	0.57	0.62	0.69	0.49	0.62	0.55

En la tabla 4, se observan los resultados al utilizar el clasificador *Bayes Net*. La mayor precisión al determinar objetividad fue de 60%, en subjetividad, la mayor

precisión fue de 76%. Las características consideradas fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 4. Resultados de experimentos utilizando n-gramas, eliminando n-gramas repetidos con el clasificador *Bayes Net.*

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.58	0.51	0.07	0.95	0.12	0.67
Bigramas, CSW, TL	0.50	0.51	0.20	0.80	0.29	0.63
Bigramas, SSW, TNL	0.55	0.51	0.06	0.95	0.11	0.66
Bigramas, SSW, TL	0.50	**0.76**	0.98	0.04	0.66	0.09
Trigramas, CSW, TNL	0.52	0.58	0.75	0.33	0.62	0.42
Trigramas, CSW, TL	0.54	0.68	0.86	0.29	0.66	0.41
Trigramas, SSW, TNL	0.53	0.60	0.78	0.33	0.63	0.42
Trigramas, SSW, TL	0.55	0.64	0.76	0.40	0.64	0.50
4-gramas, CSW, TNL	0.55	0.65	0.81	0.34	0.66	0.45
4-gramas, CSW, TL	0.57	0.69	0.81	0.41	0.67	0.51
4-gramas, SSW, TNL	0.56	0.63	0.75	0.42	0.64	0.50
4-gramas, SSW, TL	**0.60**	0.65	0.70	0.54	0.65	0.59

Posteriormente, se realizaron dos clases de experimentos. En la primera, si algún n-grama se encontraba etiquetado, tanto objetivo como subjetivo, el n-grama objetivo fue eliminado, y en la segunda, si algún n-grama se encontraba etiquetado tanto objetivo como subjetivo, se eliminó el n-grama subjetivo.

A continuación, se muestran los resultados de los experimentos al eliminar los n-gramas clasificados como objetivos.

En la tabla 5, se muestran los resultados obtenidos con *Naive Bayes*. La mayor precisión determinando objetividad es de 61%, en cuanto a subjetividad, la mayor precisión fue de 68%.

Posteriormente, en la tabla 6, se observan los resultados obtenidos con el clasificador *Bayes Net*. La mayor precisión en objetividad fue de 73%, en subjetividad, la mayor precisión fue de 69%.

Las características a considerar en ambos experimentos fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 5. Resultados de experimentos utilizando n-gramas, eliminando n-gramas objetivos repetidos con el clasificador *Naive Bayes*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.50	0.56	0.13	0.89	0.21	0.68
Bigramas, CSW, TL	0.52	0.53	0.04	0.96	0.08	0.64
Bigramas, SSW, TNL	0.50	0.56	0.12	0.90	0.19	0.69
Bigramas, SSW, TL	**0.61**	0.53	0.04	0.97	0.07	0.69
Trigramas, CSW, TNL	0.53	0.54	0.33	0.73	0.40	0.62
Trigramas, CSW, TL	0.56	**0.68**	0.78	0.43	0.66	0.52
Trigramas, SSW, TNL	0.53	0.54	0.29	0.77	0.37	0.64
Trigramas, SSW, TL	0.56	0.64	0.72	0.47	0.63	0.54
4-gramas, CSW, TNL	0.56	0.56	0.53	0.59	0.55	0.58
4-gramas, CSW, TL	0.55	0.66	0.80	0.38	0.66	0.48
4-gramas, SSW, TNL	0.55	0.57	0.55	0.57	0.55	0.57
4-gramas, SSW, TL	0.57	0.62	0.68	0.50	0.62	0.55

Tabla 6. Resultados de experimentos utilizando n-gramas, eliminando n-gramas objetivos repetidos con el clasificador *Bayes Net*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.54	0.55	0.07	0.95	0.12	0.70
Bigramas, CSW, TL	**0.73**	0.53	0.01	0.99	0.03	0.69
Bigramas, SSW, TNL	0.51	0.56	0.06	0.95	0.11	0.70
Bigramas, SSW, TL	0.66	0.52	0.00	0.99	0.01	0.69
Trigramas, CSW, TNL	0.62	0.53	0.13	0.92	0.21	0.68
Trigramas, CSW, TL	0.54	0.68	0.86	0.30	0.66	0.42
Trigramas, SSW, TNL	0.62	0.54	0.14	0.92	0.23	0.68
Trigramas, SSW, TL	0.55	0.65	0.76	0.42	0.64	0.51
4-gramas, CSW, TNL	0.55	0.63	0.78	0.37	0.64	0.47
4-gramas, CSW, TL	0.57	**0.69**	0.81	0.41	0.67	0.52
4-gramas, SSW, TNL	0.55	0.63	0.74	0.43	0.63	0.51
4-gramas, SSW, TL	0.60	0.65	0.71	0.53	0.65	0.59

A continuación, se muestran los resultados de los experimentos al eliminar los n-gramas clasificados como subjetivos:

En la tabla 7, se presentan los resultados obtenidos con el clasificador *Naive Bayes*. La mayor precisión en objetividad fue de 60% y en subjetividad fue de 67%.

Las características fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 7. Resultados de experimentos utilizando n-gramas, eliminando n-gramas subjetivos repetidos con el clasificador *Naive Bayes*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	**0.60**	0.57	0.69	0.48	0.64	0.52
Bigramas, CSW, TL	0.56	0.66	0.82	0.35	0.67	0.46
Bigramas, SSW, TNL	0.60	0.57	0.79	0.34	0.68	0.43
Bigramas, SSW, TL	0.54	0.59	0.87	0.20	0.66	0.30
Trigramas, CSW, TNL	0.55	0.54	0.61	0.48	0.58	0.51
Trigramas, CSW, TL	0.55	**0.67**	0.81	0.37	0.66	0.48
Trigramas, SSW, TNL	0.56	0.54	0.62	0.48	0.59	0.48
Trigramas, SSW, TL	0.56	0.62	0.73	0.44	0.64	0.51
4-gramas, CSW, TNL	0.56	0.56	0.60	0.52	0.58	0.54
4-gramas, CSW, TL	0.55	0.66	0.81	0.36	0.66	0.47
4-gramas, SSW, TNL	0.56	0.57	0.64	0.49	0.60	0.53
4-gramas, SSW, TL	0.57	0.61	0.69	0.49	0.62	0.55

En la tabla 8, se muestran los resultados obtenidos con el clasificador *Bayes Net Bayes*. La mayor precisión en objetividad fue de 60% y en subjetividad de 74%.

Las características fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 8. Resultados de experimentos utilizando n-gramas, eliminando n-gramas subjetivos repetidos con el clasificador *Bayes Net.*

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.54	0.70	0.98	0.05	0.70	0.09
Bigramas, CSW, TL	0.51	0.72	0.97	0.07	0.67	0.14
Bigramas, SSW, TNL	0.56	0.73	0.98	0.04	0.71	0.08
Bigramas, SSW, TL	0.52	**0.74**	0.98	0.09	0.05	0.09
Trigramas, CSW, TNL	0.54	0.64	0.87	0.23	0.67	0.34

	Precisión		Recall		Medida·F	
	O	**S**	**O**	**S**	**O**	**S**
Trigramas, CSW, TL	0.54	0.68	0.86	0.29	0.66	0.40
Trigramas, SSW, TNL	0.56	0.65	0.87	0.26	0.68	0.37
Trigramas, SSW, TL	0.56	0.63	0.76	0.41	0.64	0.50
4-gramas, CSW, TNL	0.56	0.65	0.81	0.35	0.66	0.45
4-gramas, CSW, TL	0.57	0.69	0.81	0.41	0.67	0.51
4-gramas, SSW, TNL	0.58	0.63	0.76	0.43	0.66	0.51
4-gramas, SSW, TL	**0.60**	0.65	0.70	0.54	0.65	0.59

Tabla 9. Resultados de experimentos utilizando n-gramas, eliminando n-gramas objetivos y subjetivos repetidos con el clasificador *Naive Bayes*.

	Precisión		Recall		Medida·F	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.53	0.53	0.33	0.72	0.41	0.61
Bigramas, CSW, TL	0.65	0.52	0.10	0.94	0.17	0.67
Bigramas, SSW, TNL	0.54	0.53	0.32	0.73	0.40	0.61
Bigramas, SSW, TL	**0.66**	0.57	0.39	0.80	0.49	0.67
Trigramas, CSW, TNL	0.54	0.55	0.51	0.57	0.53	0.56
Trigramas, CSW, TL	0.56	**0.68**	0.80	0.40	0.66	0.50
Trigramas, SSW, TNL	0.55	0.55	0.5	0.61	0.52	0.54
Trigramas, SSW, TL	0.56	0.63	0.73	0.45	0.63	0.53
4-gramas, CSW, TNL	0.56	0.56	0.58	0.54	0.57	0.55
4-gramas, CSW, TL	0.55	0.66	0.80	0.37	0.66	0.47
4-gramas, SSW, TNL	0.56	0.57	0.60	0.53	0.58	0.55
4-gramas, SSW, TL	0.57	0.62	0.68	0.50	0.62	0.55

Las pruebas finales, se realizaron de la siguiente manera; Si algún n-grama se encontraba etiquetado tanto objetivo como subjetivo, el n-grama fue eliminado, tanto el objetivo como el subjetivo. A continuación, se presentan los resultados:

En la tabla 9, podemos observar los resultados con el clasificador *NaiveBayes*. La mayor precisión fue 66%, en subjetividad, la mayor precisión fue 68%.

Para finalizar, en la tabla 10, podemos observar los resultados obtenidos con el clasificador *Bayes Net*. La mayor precisión en objetividad fue de 60%. En subjetividad, la mayor precisión fue de 69%.

Las características fueron: Incluyendo *stopwords* (CSW), Eliminando *stopwords* (SSW), texto sin lematizar (TNL) y texto lematizado (TL).

Tabla 10. Resultados de experimentos utilizando n-gramas, eliminando n-gramas objetivos y subjetivos repetidos con el clasificador *Bayes Net*.

	Precisión		**Recall**		**Medida·F**	
	O	**S**	**O**	**S**	**O**	**S**
Bigramas, CSW, TNL	0.59	0.52	0.07	0.95	0.13	0.67
Bigramas, CSW, TL	0.51	0.51	0.21	0.81	0.29	0.63
Bigramas, SSW, TNL	0.58	0.51	0.07	0.95	0.12	0.66
Bigramas, SSW, TL	0.50	0.75	0.97	0.07	0.66	0.13
Trigramas, CSW, TNL	0.52	0.65	0.86	0.24	0.65	0.35
Trigramas, CSW, TL	0.54	0.69	0.86	0.29	0.66	0.41
Trigramas, SSW, TNL	0.54	0.59	0.71	0.40	0.61	0.48
Trigramas, SSW, TL	0.55	0.65	0.77	0.41	0.64	0.50
4-gramas, CSW, TNL	0.55	0.65	0.81	0.34	0.66	0.45
4-gramas, CSW, TL	0.57	**0.69**	0.81	0.41	0.67	0.51
4-gramas, SSW, TNL	0.57	0.63	0.75	0.43	0.64	0.51
4-gramas, SSW, TL	**0.60**	0.65	0.70	0.54	0.65	0.59

5. Conclusiones y trabajo futuro

En este artículo se presentó un método de detección de subjetividad en noticias en español, implementando clasificadores probabilísticos, utilizando un corpus de noticias etiquetado de manera manual y clasificadores probabilísticos mencionados en la bibliografía.

Se desarrolló un corpus de noticias en español, el cual contiene notas periodísticas publicadas en México, de diversos sitios web. Este corpus sirvió como datos de entrenamiento para los clasificadores probabilísticos.

Se realizaron diversos experimentos, modificando las características de los datos del corpus, con el fin de determinar con qué clase de características y con qué clasificador se obtendría el mejor rendimiento al detectar objetividad y subjetividad. Durante la creación del corpus, se observó que no se obtuvo alguna noticia que se encontrara libre de contener oraciones subjetivas.

Se pudo observar, que en cuanto a la detección de objetividad, la mayor precisión fue de 73%, al eliminar los n-gramas objetivos repetidos, utilizando bigramas, incluyendo *stopwords* y lematizando el texto, al utilizar el clasificador *Bayes Net*. En subjetividad, la mayor precisión fue de 76%, la cual se obtuvo al utilizar el clasificador

Bayes Net, eliminando los n-gramas repetidos, utilizando bigramas sin *stopwords* y lematizando el texto.

En cuanto al trabajo futuro, se planea implementar un módulo de análisis automático basado en reglas, con el objetivo de corregir posibles errores en la clasificación automática, además de tratar de mejorar los resultados obtenidos en este trabajo. Además, se planea tratar de identificar qué porcentaje de oraciones subjetivas puede contener una noticia para ser considerada como objetiva.

Referencias

1. Estudio sobre los hábitos de los usuarios de Internet en México. Asociación Mexicana de Internet. https://www.amipci.org.mx/es/estudios
2. Gómez, R., Sosa-Plata, G., Bravo, Téllez-Girón, P., Dragomir, M., Thompson, M.: Los medios digitales: México. Open Society Foundations. http://www.opensocietyfoundations.org
3. Wiebe, J. M.: Tracking point of view in narrative. Computational Linguistics, pp. 233-287 (1994)
4. Wiebe, J., Bruce, R., Martin, M., Wilson, T., Bell, M.: Learning subjective languaje. Computational Linguistics, pp. 277–308 (2004)
5. Salaverría, R., Cores, R. .: Géneros periodísticos en los cibermedios hispanos (2005)
6. Abundis, F.: Los medios de comunicación en México. Parametría, Investigación de opinión y mercados (2006)
7. Panicheva, P., Cardiff, J., Rosso, P.: Identifying Subjective Statements in News Titles Using a Personal Sense Annotation Framework. In: American Society for Information Science and Technology, pp. 1411–1422 (2013)
8. Sukhum, K., Nitsuwat, S., Choochart, H.: Opinion Detection in Thai Political News Columns Based on Subjectivity Analysis. In: 7th International Conference on Computing and Information Technology (2011)
9. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, C., Riloff, E., Patwardhan, S.: OpinionFinder, A system for subjectivity analysis. In: Conference on Empirical Methods in Natural Languaje Processing, pp. 34–35 (2005)
10. Lin, C., He, Y., Everson, R.: Sentence subjectivity detection with weakly-supervised learning. In 5th International Joint Conference on Natural Languaje Processing, pp. 1153–1161 (2011)
11. Bruce, R. F., Wiebe, J. M.: Recognizing subjectivity; A case of study of manual tagging. Natural Languaje Engineering, pp. 1–16 (1999)
12. Bouckaert, R. R., Eibe, F., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: Weka manual for version 3-7-11, Waikato (2014)

13. Wang, S., Manning, C. D.: Baselines and Bigrams, Simple, good sentiment and topic classification. In: 50th Annual Meeting of the Association for Computaniotal Linguistics, pp. 90–94 (2012)
14. McCallum, A., Nigam, K.: A comparison of events model for Naive Bayes Text classification. In: Learning for text categorization (1998)