Detección automática de noticias falsas

Andy González Peña
a.gonzalez@estudiantes.matcom.uh.cu

Matemática y Computación, Universidad de la Habana, 2018

Abstract. Con el auge de las redes sociales el consumo de información negativa ha aumentado exponencialmente al punto que se requieren de nuevos mecanismos que, al menos, regulen su circulación. Encabezando la lista se encuentran las noticias falsas que, a pesar de siempre haber existido ya sea por fines políticos u otros, toman nueva relevancia al aparentar pertenecer a canales oficiales o aparecer en la red de confianza de un usuario de una determinada red social. En este trabajo se presentan las características principales del problema antes de la aplicación en la web que se conoce en la actualidad, es decir, se propone como reconocer patrones de una noticia desde el punto de vista estático en el que no se puede hacer nada por no recibir tal noticia, todo lo contrario, se recibe y se decidirá la validez a partir de un análisis utilizando técnicas de *Machine Learning* (ML).

1 Introducción

Muchos esfuerzos son dedicados en la actualidad para enfrentar la sombra que proyectan las divulgaciones falsas y negativas sobre el sano uso de las redes sociales. En este trabajo se resumen las principales técnicas existentes para resolver esta problemática así como las diferencias entre los resultados aplicados a modelos que solamente dependen de la noticia en cuestión como las propuestas enfocadas a las redes sociales, esto se puede encontrar en la sección de **Estado del Arte**.

Tomando en cuenta el análisis arriba mencionado se procederá a proponer una solución, en la sección de igual nombre, basada en técnicas de ML con datos como titular y texto exclusivamente y utilizando un *corpus* de clasificación en solamente dos categorías, reales o falsas, utilizando las principales características del procesamiento de lenguaje natural (NLP) que no dependerá de entidades externas como en una plataforma social, es decir, autores o referencias, sino utilizando únicamente los patrones más abundantes dentro del texto. En consecuencia, se ofrecerán los resultados alcanzados y la relevancia de los mismos en comparación con los sistemas ya existentes en explotación.

Dentro de la rama de minería de datos, la detección automática de noticias falsas se encuentra aún en sus inicios y existen muchos trabajos y discusiones abiertas, por tanto, la propuesta de solución no deberá ser tomada a la ligera ni con el afán de implementación óptima. De manera concluyente se presentan aspectos teóricos que podrían brindar mejores resultados así como futuras incorporaciones que deberán ser aplicadas con el objetivo de proponer un sistema altamente certero.

Con este trabajo se persigue la consolidación de los conocimientos básicos de minería de datos, inteligencia artificial y sistemas de información en cuestión, la aplicación de esquemas de procesamiento de lenguaje natural (NLP) y abrir nuevos debates constructivos que impacten en un grado positivo sistemas de estas características para, con ello, poder disfrutar en mayor medida de la parte sana de la actual difusión de la informarción.

Sin pretensiones de alta profundización en diversas temáticas especialmente aplicadas a las redes sociales, o la propia definición de noticias falsas, como pueden ser factores psicológicos, políticos, sociales u otros, se procede

a continuar con la siguiente sección, ceñida a elementos puramente computacionales.

2 Estado del Arte

A pesar de ser una temática relativamente nueva, en consecuencia de relevancia y de lo que apremia como problema social, existen numerosos trabajos que investigan e incluso ya explotan los bases del conocimiento adquirido. Sitios reportan hasta 90 % de efectividad para clasificar una noticia en verdadera o falsa y para ello se bastan de las técnicas de NLP y ML, sin embargo utilizan otros datos relevantes a su entorno, datos sociales como pueden ser fuentes y referencias, o grado de pertenencia a grupos de confiabilidad en cuanto información, entre otras.

Los sistemas que indican un menor porcentage de efectividad analizan puramente texto, y hasta el momento la cifra más elevada se encuentra en 76 %. Parecería una cifra no tan alta, sin embargo, estudios indican que los humanos se equivocan el 70 % de los casos en definir la veracidad de una noticia. Esto significa que para ser un sistema que automatiza el trabajo exhaustivo de analistas y que propone mejores resultados que ellos no está nada mal.

2.1 Fundamentos NLP

Se define la problemática como una situación específica de ML, se tiene una base de conocimientos dados por noticias y sus clasificaciones y se pretende diseñar un modelo que aprenda de ello y sea capaz, entonces, de predecir nuevos resultados a partir de lo que ya sabe. La ciencia de los datos se divide en dos partes cuando acontece un problema de este tipo: primeramente se conocen los modelos existentes de ML para afrontarlo, pero sus entradas son vectores de datos y la base de conocimientos solamente posee texto, es precisamente esa la segunda cuestión y quizás la más importante del problema, la extracción de datos relevantes para que el modelo propuesto clasifique correctamente.

Entonces, descartando los datos que podrían ser relevantes para los sistemas de manera general, quedamos solamente con los que podría ofrecer un texto dentro de la base del conocimiento, así se definen los patrones linguisticos del lenguaje natural, los cuales se dividen en tres grupos fundamentales: lexicográficos, sintácticos y semánticos. Los elementos lexicográficos pueden ser, por ejemplo, una tabla de frecuencias de palabras con lo que se pueden denotar estilos en los usos de palabras dentro del texto. Los elementos sintácticos podrían ser una tabla de frecuencias de frases u oraciones gramaticales con lo que se pueden denotar secciones relevantes a la noticia en su completitud. Por último, tenemos los elementos semánticos del texto, estos podrían tomar la forma de similaridad espacio-vectorial entre oraciones.

Los problemas que NLP contiene son mucho más que lo que aborda este trabajo, sin embargo en cuanto a la extracción de los elementos antes mencionados existen ya muchas técnicas. La tokenización por oraciones y por palabras son soluciones que hoy podemos llamar *straight-forward* y son claves para el resto del análisis. Para la parte sintáctica, la herramienta más común es la utilización de *POS-tagging*, basado en la implementación de un perceptrón pre-entrenado de la librería de su pertenencia para reconocer *Parts-of-Speech* o partes del texto apoyándose en una gramática para extraer las que se consideran relevantes al problema. En este tema existen grandes debates acerca de que parte del texto representa en mayor grado su oración o su texto propiamente. Para la extracción de elementos semánticos la práctica más común está basada en la similaridad espacio-vectorial la cual, a su vez, se apoya en la técnicas como esquema de N-gramas o *Bag-of-Words*. Estas técnicas vectorizan y cuantifican la relación de una frase de tamaño n, o un conjunto de palabras, en su ámbito oracional.

2.2 Fundamentos Aplicados ML

El modelo de ML a utilizar es la otra sección del diseño, sin embargo, cual sea el escogido, todos los vectores y matrices recogidos utilizando las técnicas vistas anteriormente deben ser nor-

malizados y recontorneados para acomodarse al modelo; por ejemplo, si se tenía como resultado una matriz y se requiere de un conjunto de datos plano, entonces la primera será aplanada, es decir, convertida a un vector.

La práctica común para la detección de noticias falsas utiliza los algoritmos de redes neuronales para el entrenamiento o aprendizaje del modelo de clasificación en falsas o verdaderas. Sin embargo, las categorías oficiales de estos sistemas son: basados en conocimientos y basados en estilos. Los primeros se acercan mucho a la problemática en la web y como enfocarlo desde tal punto de vista, mientras que la detección de estilos propone un profundo análisis textual para definir la veracidad de la noticia. Para la detección de estilos se han utilizado, incluso, redes convolucionales dependiendo de la magnitud de datos a procesar.

3 Propuesta de Solución

Dadas las características del *corpus* a utilizar, el cual contiene título, cuerpo y clasificación, se propone la implementación de un modelo basado en estilos para insertar profundidad en el análisis textual de la base de conocimientos. En consecuencia, el procedimiento de ML se llevará aplicando las secuencias de entrenamiento, validación y predicción que conlleva el uso de los algoritmos de redes neuronales, sin llegar al despliegue total de las redes convolucionales. Junto a ello, para la extracción de *features* se procede a la implementación de los aspectos antes mencionados de NLP, es decir, el análisis lexicográfico, sintáctico y semántico.

3.1 Especificidades y Dependencias

La solución se ha escrito en Python (v. 3.6.1) de la distribución de Anaconda y se utilizan diferentes librerías para cumplir los objetivos planteados. Para la implementación de redes neuronales, se utiliza la librería de *Tensorflow* y dentro, específicamente, el *framework* de *Keras* el cual provee un mayor nivel de abstracción en la resolución modular de los problemas de ML. Para los análisis de patrones linguísticos se utiliza NLTK, herramienta poderosa que contiene muchas sub-librerías de alto valor funcional, específicamente y aplicado a este trabajo, se denotan las secciones de tokenización, de POS-tagging y extracción gramatical. Mientras, para la vectorización en el cálculo de similaridad, se utiliza la librería de *sklearn*, en las sección de extracción de *features*.

3.2 Estructura y Funcionalidades

Se comienza con la carga de la base de conocimientos y se procede a el análisis de patrones existentes en la misma. Dado que el proyecto va enfocado al procesamiento de textos, continúa a la extracción de características puntuales o de estilos dentro de los textos de noticias, para este caso se denominan los *features* linguísticos.

Dentro de las características linguísticas, el análisis se divide en tres: lexicográfico, sintáctico y semántico. En el primero se pretende encontrar el uso o frecuencias de palabras dentro de el texto de la noticia, y con ello proponer la primera característica de estilos: probabilidad de que una noticia sea falsa si el estilo de repetición de palabras contiene unas frecuencias de x_1, x_2, ..., x_n. De la misma manera, el análisis sintáctico pretende encontrar las repeticiones de frases que se asumen que tienen peso dentro de cada oración del texto como pueden ser las nominaciones (sustantivos) con sus calificaciones (adjetivos) junto al posible uso de artículos, entre otros; así se construye otra tabla de frecuencias y se propone otra característica de estilos: probabilidad de que una noticia sea falsa si el estilo de repetición de frases contiene unas frecuencias de y_1, y_2, ..., y_n. Por último, el análisis semántico procede al cálculo de la similaridad entre oraciones, es decir, se analiza las relaciones de concordancia en el texto, entonces se obtiene la probabilidad de que una noticia sea falsa dado que su estilo contiene tales grados de concordancia a la idea general que indica, esto se calcula mediante combinaciones de frases dentro de las mismas oraciones según el modelo de N-gramas, frecuencias y sus inversas (tf*idf) para denotar su peso en la idea y la normalización matricial para computar el grado de similaridad.

Una vez extraídos los principales patrones se obtienen de ellos matrices de datos relevantes para la clasificación, sin embargo, estas pueden ser de tamaños variables y, por ende, es necesario llevarlas todas a la misma medida de tal forma que se conforme la descripción vectorial de una noticia. Una vez computado el vector de cada patrón se continúa a la combinación de ellos mediante la concatenación de listas.

Estos vectores conformarán los verdaderos conjuntos de entrenamiento y validación para las redes neuronales. Con el objetivo de lograr esto, se utiliza el método de *k-fold*, específicamente *2-fold* en el que se divide el *corpus* en dos conjuntos: los llamados entrenamiento y validación. En el primero se ajusta el modelo a minimizar la función objetivo que describe el conjunto de datos y la segunda se dedica a evaluar el índice de correctitud de las predicciones del sistema.

4 Resultados

Se utilizan las métricas de precisión, recobrado (*recall*), F1 y eficacia (*accuracy*). Se define las siguientes variables:

- TP (*True Positive*) : cuando se predice una noticia falsa y está anotada realmente como una noticia falsa.
- TN (*True Negative*) : cuando se predice una noticia verdadera y está anotada realmente como una noticia verdadera.
- FN (*False Negative*) : cuando se predice una noticia verdadera y está anotada realmente como una noticia falsa.
- FP (*False Positive*) : cuando se predice una noticia falsa y está anotada realmente como una noticia verdadera.

Luego se formulan las métricas según ellas de la siguiente manera:

$$\text{Precisión} = \frac{|TP|}{|TP|+|FP|}$$

$$\text{Recobrado} = \frac{|TP|}{|TP|+|FN|}$$

$$\text{F1} = 2 * \frac{Precision * Recobrado}{Precision + Recobrado}$$

$$\text{Eficacia} = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$$

Sin embargo los porcentajes antes mencionados son valores de eficacia del sistema. Aunque no se presume de competir con resultados altos según indica el estado del arte, para sistemas con estas características se tienen 76 %, se ha obtenido buenos valores de eficacia.

A continuación se presentan ejemplos de corridas del modelo para mostrar sus evaluaciones según las métricas indicadas:

Medidas %	**Léx.**	**Sint.**	**Sem.**	**Todos**
Precisión	53	62	78	72
Recobrado	63	36	53	70
F1	58	46	64	71
Eficacia	54	57	70	72

Como se puede observar en la tabla, con las primeras tres medidas corridas se obtiene un alto desbalance entre las métricas de precisión y recobrado, aumentando poco a poco los valores de eficacia total hasta que en la tercera se proveen de medidas cercanas al objetivo a perseguir; sin embargo, es la combinación de todas las cuales aumentan esa eficacia total requerida hasta llegar a un 72 %, aún por encima de los índices de reconocimiento humano y solamente un 4 % por debajo del mejor sistema actual en explotación.

5 Conclusiones

A pesar de haber obtenido buenos resultados quedará pendiente el mencionado 4 %. Para ello se propone como posible solución y futura discusión la aplicación de algoritmos de *clustering* para la agrupación por temas y con ello proveer de una nueva característica descriptiva: calcular la probabilidad de que, dado un tema asociado, la noticia sea falsa.

De esta forma se asume la solución como aún no óptima denotando varias características posibles que podrían ser utilizadas para obtener resultados más altos. Obviamente, la completitud de esta idea se encuentra todavía en estudios. Por estas razones, el tema se presenta tan atractivo ya que propone un desafío para las técnicas actuales.

Bibliography

Blázquez-Ochando, M. (2018). El problema de las noticias falsas: detección y contramedidas.

Dietterich, T. Ensemble methods in machine learning. multiple classifier systems, first int. In *Workshop, MCS2000, Cagliari, Italy*, pages 1–15.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10.

Garg, A. and Roth, D. (2001). Understanding probabilistic classifiers. In *European Conference on Machine Learning*, pages 179–191. Springer.

Gish, H. (1990). A probabilistic approach to the understanding and training of neural network classifiers. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 1361–1364. IEEE.

Horne, B. D. and Adali, S. (2017). This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer.

Riedel, B., Augenstein, I., Spithourakis, G. P., and Riedel, S. (2017). A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.