ISSN 1870-4069

A Sentiment Analysis Approach for Drug Reviews in Spanish

Karina Castro Pérez[1], José Luis Sánchez Cervantes[2], María del Pilar Salas Zárate[1],
Luis Ángel Reyes Hernández[1], Lisbeth Rodríguez Mazahua[1]

[1] Tecnológico Nacional de México,
Instituto Tecnológico Orizaba,
Mexico

[2] CONACYT,
Instituto Tecnológico de Orizaba
Mexico

karinacastro.058@gmail.com, jlsanchez@conacyt.mx,
{msalasz, lreyes, lrodriguezm}@ito-depi.edu.mx

Abstract. The analysis of opinions in the medical context is of great relevance for health care since it allows us to gain insight from the experiences and opinions of patients and health professionals regarding nutrition, exercise, and other health related issues. The rise in the application of opinion mining in recent years is a direct consequence of the growth of social networks and blogs that generate a large volume of unstructured data, however, the manual review of such data is not feasible due to the amount that is generated in real time. Thus, opinion summarization systems that use Web Scraping techniques and opinion mining are needed. In this sense, this work presents a solution proposal under a hybrid approach based on both semantic approach and machine learning approach for the development of an opinion mining analysis system applying Web Scraping and Natural Language Processing (NLP) techniques to know the users' experiences about drugs for chronic-degenerative diseases available in blogs, video blogs and specialized websites in Spanish language.

Keywords: Web scraping, natural language processing, opinion mining, Spanish language analysis, drugs opinion.

1 Introduction

Opinion mining is an area of great importance for the coarse application that has, focuses on analyzing opinions, sentiments, evaluations, assessments, attitudes and emotions of people towards entities such as products, services, organizations, individuals, problems and events. [1]; the application of this field has increased over the years as a result of the growth of social networks and blogs that generate vast volume amounts of unstructured data, nevertheless, the manual revision of the data is not feasible due to the amount that is generated in real time.

Therefore, it is necessary to apply Web Scraping and opinion mining techniques that allow summarizing the information and obtaining precise knowledge, of interest in a certain area, which will result useful in a decision making process.

According to the World Health Organization (WHO) in 2016 [2], diabetes mellitus, hypertension, cardiovascular diseases, cancer, among others diseases catalogued as chronic-degenerative, are positioned among the top ten causes of mortality in Mexico and worldwide; as a result, health systems need to analyze important aspects such as eating habits, exercise and treatments. In this context, the application of opinion mining is valuable because it allows analyzing the comments of patients and health professionals to identify symptoms and drugs related to chronic-degenerative diseases. In addition, it allows health specialists to speed up the process of identification and selection of the drugs that they prescribe, allowing them to dedicate more time to the physical exploration of the patient to prevent additional complications to the disease, which results in higher quality attention for the patient. The main contribution of this work is a hybrid approach that applies semantics through a tagged corpus and supervised machine learning for an opinion analysis system for drugs, searching blogs, video blogs and specialized websites, in the Spanish language, implementing Web Scraping techniques and Natural Language Processing (NLP).

This paper is structured as follows: Section 2 presents a set of recent works related to our proposal. It describes initiatives of healthcare-oriented opinion mining and a comparison of articles that address polarity detection as an opinion mining activity, the use of NLP and the sources from which the data were obtained. The proposed hybrid approach for the opinion analysis system for drugs in Spanish based on Web Scraping techniques is presented in Section 3; Section 4 presents a case study for polarity analysis in comments published in Spanish language on drugs for chronic-degenerative diseases. Finally, the Section 5 presents the conclusions and future work.

2 Related Work

An analysis that describes the difficulty of finding Adverse Drug Events (ADE) in clinical trials conducted by pharmaceutical organizations was developed by Lee et al. [3]. For this reason, authors examined deep learning models, however, it was found that employing such a method is costly given the scarcity of Twitter ADE tweets. Similarly, a method to extract ADE of Twitter, through a five-step channel was proposed in [4]. Such steps are: 1) Tweet capture; 2) Data pre-processing; 3) Drug-related classification; 4) Tweet sentiment analysis and, 5) ADE extraction from Twitter data. The potential of sentiment analysis in medicine with data from clinical narratives and medical social networks was described by Denecke and Deng [5].

Through a literature review, we summarize linguistic peculiarities of sentiment in medical texts. On the other hand, in [6], an investigation of the NLP techniques for the extraction of opinions and sentiment analysis was carried out, which identified the stages of pre-processing required to structure the texts. It was found that tokenization is an essential task for Chinese and Japanese, among others, because their words are composed differently from the Latin alphabet.

Table 1. Comparative of related works.

Initiative	Approach	Polarity detection	Data Source	NLP	Language
Lee et al. [3]	Semi-supervised convolutional neural network.	✓	Twitter	X	English
Y. Peng et al. [4]	Supervised machine learning and linguistic method.	✓	Twitter	✓	English
Denecke & Deng [5]	-----	✓	Clinical narratives, social networks and specialized websites.	X	English
Solangi et al. [6]	-----	✓	-----	✓	English, Chinese, Japanese, Arabic, French, Spanish, and German
Salas-Zárate et al. [7]	Semantic	✓	Twitter	✓	English

A method for sentiment analysis that detects diabetes-related aspects in tweets, using ontologies to semantically describe the relationships between concepts in the specific domain. Salas-Zárate et al. [7] presented a proposed sentiment classification approach divided into three main components: the pre-processing module for cleaning and correcting text, the semantic annotation module for aspect detection, and the sentiment classification module that calculates polarity.

As shown in Table 1, the analyzed works use one of several approaches, such as the application of algorithms for the classification of sentiment and the use of semantic methods, both of which are used in our system as part of a hybrid approach. Unlike [3] and [8], our approach gets comments from a variety of sources such as forums, blogs and video blogs, where there is relevant information that has not been analyzed in detail, so the scarcity of comments does not prove to be an obstacle as it is with getting tweets about medication from the social network Twitter.

Our approach makes a review on chronic-degenerative diseases, so a vast amount of comments is obtained through Web Scraping, and they contain information about sentiments regarding a drug, the dosage of the drug, the price and even the adverse effects that it has on patients.

On the other hand, a hybrid approach is proposed in [4] using sentiment classification algorithms such as Syntactic dependency paths, in addition, linguistic method through external tools, unlike our approach, we adopt the supervised machine learning that integrates a semi-automatic tagged corpus that makes use of a dictionary of positive and negative words to tag the corpus, to ensure that the corpus is correct,

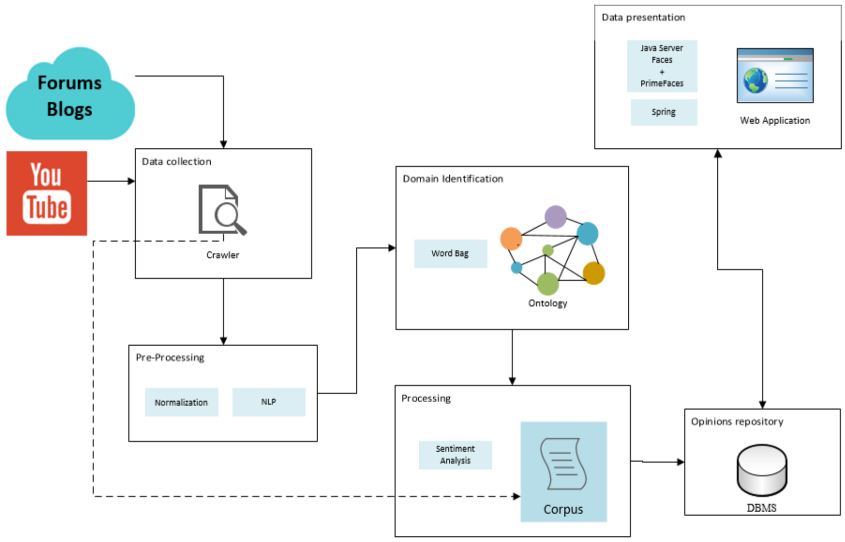


Fig. 1. Architecture of *SentiScrap* system.

two health specialists analyzed the comments collected and tagged, where they found mentions of diseases, medications and symptoms and the use of automatic learning. In addition, our approach gets first-hand comments from patients unlike the clinical narratives discussed in [5] which are subject to interpretation by third parties, by doctors and nurses, so they are less accurate. Finally, as mentioned in [7], very scarce work has been done in the implementation of opinion mining and NLP in languages other than English, such as Chinese and Japanese, among others, because it requires a great effort for implementation and analysis, while our approach addresses opinion mining and NLP for comments in the Spanish language as it is recognized as being the second most spoken in the world.

3 Approach

In order to achieve a successful outcome, a hybrid approach using semantics and automatic learning is proposed for a Spanish opinion analysis system for chronic-degenerative disease drugs, based on Web Scraping techniques, consisting of six main modules and a corpus: 1) Data collection module; 2) Pre-processing module; 3) Domain identification module; 4) Processing module; 5 Opinions repository, and 6) Data presentation. Figure 1 shows the architecture of system. In this approach, a corpus was created with 280 comments obtained from forums on diseases such as type 2 diabetes mellitus, hepatitis and hypertension. Given that polarity detection is limited in the Spanish language due to the lack of data sets, we made a corpus tagged semi-automatically with the use of a dictionary that was manually built, from the analysis of comments found in specialized forums, blogs and video blogs to identify the words that tell a positive comment from a negative one. To ensure the effectiveness of corpus labeling, two health specialists were consulted; they manually reviewed each comment to corroborate that both, the positive and negative comments presented the corresponding tag.

3.1 Data Collection Module

In this module, we collected reviews on drugs and symptoms of chronic-degenerative diseases from forums, blogs, and video blogs. Specifically, diabetes, hypertension, and hepatitis diseases were considered.

3.2 Pre-Processing Module

The pre-processing of data is an important step for the normalization of the text, therefore, we chose to use three phases for the treatment of our data:

1. Delete unusual characters: comments contain special characters that do not provide information, so they are deleted.
2. Delete duplicate comments: This step is important, because duplicate comments affect the final result of the analysis, therefore it is important to ensure that duplicate comments are removed.
3. Delete comments that only have URLs: comments that only include links to other sites do not contribute as a comment to analyze polarity, for this reason they are discarded.

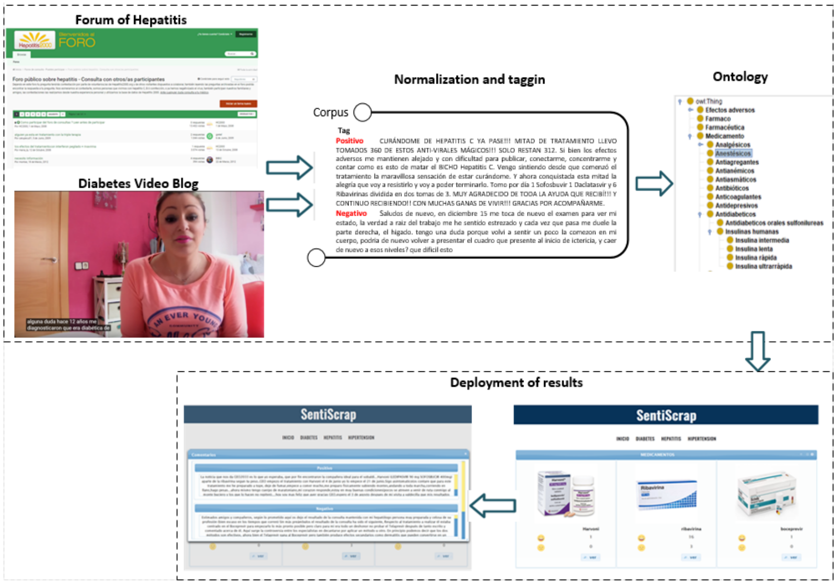
The application of these tasks on the comments ensures a better analysis of sentiments, however, the incorrect use of language is a common scenario, created by the use of abbreviations or spelling errors on behalf of the users, requiring a greater effort to carry out opinion mining activities, for that reason, this module also makes use of a spell checker.

3.3 Domain Identification Module

The identification of words has high relevance because it allows the verification that the comments obtained is the mention of a drug prescribed for diabetes, hypertension and hepatitis, this provides as a result a set of data with valuable information for analysis. Therefore, a bag of words is used, made through an ontology of medical domain based on Snomed [8].

3.4 Processing Module

This module adopts the supervised automatic learning approach which makes use of a semi-automatic labeled corpus, necessary to train the algorithm that performs the sentiment analysis, this permits it to recognize new opinions in the Spanish language and to classify them correctly.



Positivo CURÁNDOME DE HEPATITIS C YA PASE!!! MITAD DE TRATAMIENTO LLEVO TOMADOS 360 DE ESTOS ANTI-VIRALES MÁGICOS!!! SOLO RESTAN 312. Si bien los efectos adversos me mantienen alejado y con dificultad para publicar, conectarme, concentrarme y contar como es esto de matar el BICHO Hepatitis C. Vengo sintiendo desde que comenzó el tratamiento la maravillosa sensación de estar curándome. Y ahora conquistada esta mitad la alegría que voy a resistirlo y voy a poder terminarlo. Tomo por día 1 Sofosbuvir 1 Daclatasvir y 6 Ribavirinas dividida en dos tomas de 3. MUY AGRADECIDO DE TODA LA AYUDA QUE RECIBÍ!!! Y CONTINUO RECIBIENDO!! CON MUCHAS GANAS DE VIVIR!!! GRACIAS POR ACOMPAÑARME.

Negativo Saludos de nuevo, en diciembre 15 me toca de nuevo el examen para ver mi estado, la verdad a raiz del trabajo me he sentido estrezado y cada vez que pasa me duele la parte derecha, el higado. tengo una duda porque volvi a sentir un poco la comezon en mi cuerpo, podria de nuevo volver a presentar el cuadro que presente al inicio de ictericia, y caer de nuevo a esos niveles? que dificil esto

Fig. 2. *SentiScrap* Workflow.

3.5 Opinions Repository Module

The opinions and polarity of the drugs resulting from the analysis are stored in a database to keep the data available for consultation.

3.6 Data Presentation Module

A web interface is presented that supports the interaction of health specialists with the "*SentiScrap*" system to know the comments and polarity of the drugs prescribed for chronic-degenerative diseases, specifically diabetes mellitus, hepatitis and hypertension.

4 Case Study: Polarity Analysis in Comments Published in Spanish Language on Drugs for Chronic-Degenerative Diseases

For this case study, the medical domain of application was selected, specific for chronic-degenerative diseases. Suppose that a health specialist needs to know quickly the opinion of patients undergoing treatment to alleviate hepatitis C disease:

- How will the specialist get patient feedback in Web forums in Spanish language?
- How will the health specialist know which comments have a positive or negative impact?

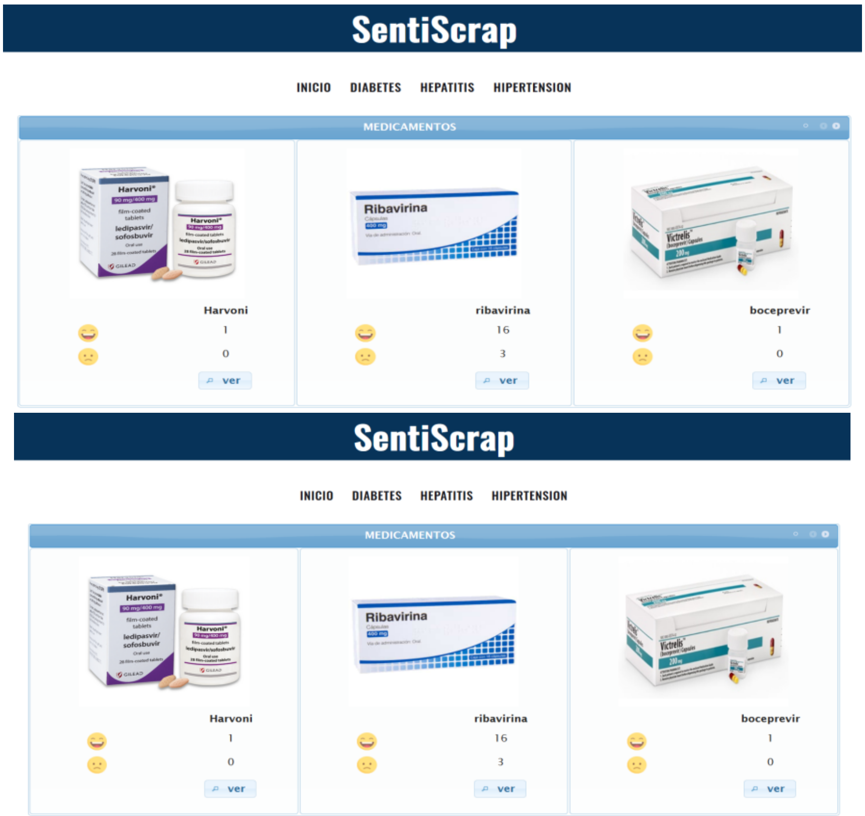


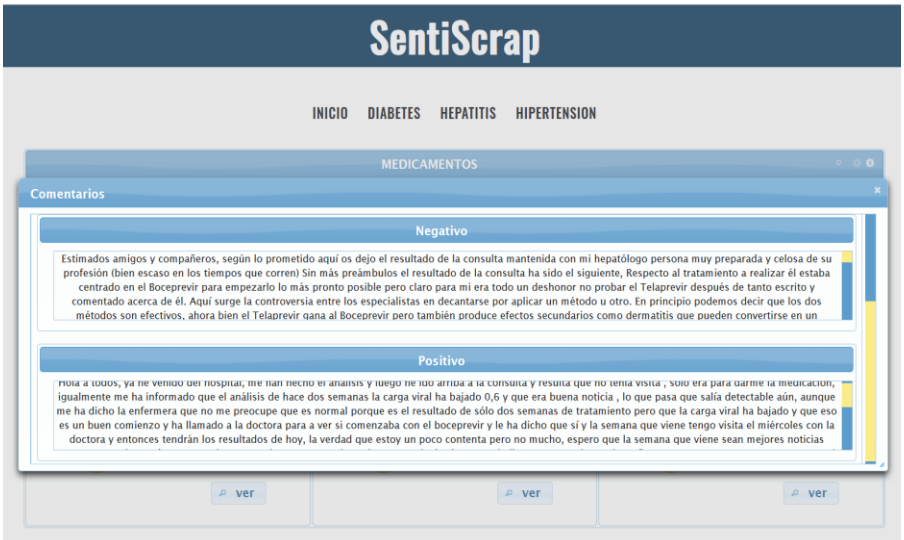
Fig 3. *Hepatitis* options of the *SentiScrap* system

- How will the specialist be able to identify the best medications prescribed for the treatment of hepatitis C?

As a solution, it is intended that the health specialist has access to the information analyzed through the "*SentiScrap*" system. Figure 2 describes the workflow for user interaction with the system's functionalities.

Suppose that doctor wants to know the prescription drugs for hepatitis C, those that patients comment on in the forums. The user selects the word "Hepatitis" in the menu options, the system generates a query to show the medicines associated to the disease, and the number of positive and negative comments of each of these drugs. Figure 3 shows the results of the polarity analysis on three prescription drugs for hepatitis, Harvoni, Ribavirina and Boceprevir. The polarity reveals that Ribavirina has 16 positive and 3 negative opinions, while Harvoni and Boceprevir have the same number of positive and negative opinions. These results, provide information to the doctor to know the drug that receives better opinions from users.

In addition, the user would like to know the comments on each drug, which, by clicking on the "ver" button, this action opens a modal window with each comment



Estimados amigos y compañeros, según lo prometido aquí os dejo el resultado de la consulta mantenida con mi hepatólogo persona muy preparada y celosa de su profesión (bien escaso en los tiempos que corren) Sin más preámbulos el resultado de la consulta ha sido el siguiente, Respecto al tratamiento a realizar él estaba centrado en el Boceprevir para empezarlo lo más pronto posible pero claro para mi era todo un deshonor no probar el Telaprevir después de tanto escrito y comentado acerca de él. Aquí surge la controversia entre los especialistas en decantarse por aplicar un método u otro. En principio podemos decir que los dos métodos son efectivos, ahora bien el Telaprevir gana al Boceprevir pero también produce efectos secundarios como dermatitis que pueden convertirse en un

Hola a todos, ya he venido del hospital, me han hecho el análisis y luego he ido arriba a la consulta y resulta que no tenía visita , sólo era para darme la medicación, igualmente me ha informado que el análisis de hace dos semanas la carga viral ha bajado 0,6 y que era buena noticia , lo que pasa que salía detectable aún, aunque me ha dicho la enfermera que no me preocupe que es normal porque es el resultado de sólo dos semanas de tratamiento pero que la carga viral ha bajado y que eso es un buen comienzo y ha llamado a la doctora para a ver si comenzaba con el boceprevir y le ha dicho que sí y la semana que viene tengo visita el miércoles con la doctora y entonces tendrán los resultados de hoy, la verdad que estoy un poco contenta pero no mucho, espero que la semana que viene sean mejores noticias

Fig 4. *SentiScrap* comments modal window

referring to the drug and its classification with respect to whether it is a positive comment or a negative one. Figure 4 shows a modal window with the collected comments.

The proposed solution aims to show that the information contained in the forums and blogs is of great relevance to health specialists, because by accessing the comments of patients, the doctor can know the experience of each patient, thus consulting first-hand data, which are useful for decision making.

5 Conclusions and Future Work

The analysis of the related work shows the existence of great opportunity in the application of studies related to the analysis of opinions and polarity detection in specialized websites, blogs or video blogs. For this reason, we propose a hybrid approach, through supervised machine learning and the use of semantics through a tagged corpus. The approach allows the analysis of the sentiments of the opinions for the drugs prescribed for chronic-degenerative diseases in a successful way, reducing time and effort in the search for relevant information on diabetes, hypertension and hepatitis diseases.

As future work, it is contemplated to incorporate more aspects to the opinions analysis, such as the adverse effects to the medicines, time of the treatment and price, likewise, we add the information to the web application, through graphs for a better visualization and analysis of the information, allowing to improve the taking of decisions that the specialists of the health make. In addition, we consider adding a module to validate users' opinions about medicines. This validation consists to rely of experts in the medical domain, such as cardiologists, nephrologists, among others, verifying that the opinions made by users are adequate, mainly to avoid self-medication.

Acknowledgments. This research work was sponsored by the National Council for Science and Technology (CONACYT) and the Secretariat of Public Education (SEP). The authors are grateful to Tecnológico Nacional de México (TNM) for supporting this work.

References

1. Liu, B.: Sentiment Analysis and Opinion Mining, pp. 168 (2012)
2. WHO: Las 10 principales causas de defunción. https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death (2019)
3. Lee, K., Qadir, A., Hasan, S.A., Datla, V., Prakash, A., Liu, J., Farri, O.: Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In: Proceedings of the 26th International Conference on World Wide Web, pp. 705–714 (2017)
4. Peng, Y., Moh, M., Moh, T.S.: Efficient adverse drug event extraction using twitter sentiment analysis. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1011–1018 (2016)
5. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine, 64(1), pp. 17–27 (2015)
6. Solangi, Y.A., Solangi, Z.A., Aarain, S., Abro, A.G., Mallah, A., Shah, A.: Review on Natural Language Processing (NLP) and its toolkits for opinion mining and sentiment analysis. In: IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1–4 (2018)
7. Salas-Zárate, M. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M.Á., Valencia-García, R.: Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational Mathematical Methods in Medicine, 2017(5), pp. 1–9, (2017)
8. NCBO BioPortal. https://bioportal.bioontology.org/ontologies (2019)