CLASIFICACIÓN DE REPORTES CLÍNICOS PARA APOYAR EL DIAGNÓSTICO DEL CÁNCER

José Alejandro Reyes Ortiz
Universidad Autónoma Metropolitana
jaro@correo.azc.uam.mx

Beatriz Adriana González Beltrán
Universidad Autónoma Metropolitana
bgonzalez@correo.azc.uam.mx

Mireya Tovar Vidal
Benemérita Universidad Autónoma de Puebla
mtovar@cs.buap.mx

Resumen

El procesamiento automático de textos clínicos ha tomado relevancia en los últimos años, debido a que, diariamente, se genera una gran cantidad de información electrónica que no está estructurada. Este procesamiento puede apoyar a la toma de decisiones clínicas para establecer un tratamiento o realizar un diagnóstico. Este artículo presenta un enfoque de clasificación supervisada de reportes clínicos mediante el algoritmo de Máquinas de Soporte Vectorial (MSV). Se utiliza información lingüística de los textos, con la finalidad de apoyar el diagnóstico de cuatro tipos de cáncer: estómago, pulmonar, cáncer de pecho y cáncer de piel. Una evaluación de información lingüística como el uso de verbos, sustantivos y adjetivos fue desempeñada sobre el conjunto de reportes clínicos. Los resultados de la evaluación de nuestro enfoque son prometedores y proporcionan un referente como herramienta para el procesamiento de textos clínicos en apoyo a los diagnósticos clínicos.

Palabras Claves: Apoyo al diagnóstico de cáncer, características lingüísticas, clasificación de textos, procesamiento de lenguaje natural.

Abstract

Automatic processing of clinical texts has become relevant in recent years, due to the large amount of electronic and unstructured data that is produced daily. This processing can support clinical decision making such as establishing a treatment or providing a diagnosis. This paper presents a supervised classification of clinical reports using the Support Vector Machine (SVM) algorithm and linguistic information from texts, in order to support the diagnosis of four types of cancer: digestive cancer, lung cancer, breast cancer and skin cancer. An evaluation of linguistic information such as the use of verbs, nouns and adjectives was performed. Evaluation results of our approach are promising and serve as a reference to the processing of clinical texts as support for clinical diagnoses.

Keywords**:** *Cancer diagnosis support, linguistic features, natural language processing, text classification.*

1. Introducción

En la actualidad, en el dominio clínico, se generan grandes cantidades de textos o reportes clínicos expresados en lenguaje natural (datos no estructurados), tales como: notas pre-operatorias, notas de altas, reportes radiológicos, reportes de exámenes y hallazgos, notas de admisión, entre otros. El procesamiento automático de esta información es complejo y costoso, ya que no tiene una estructura semántica y procesable por computadoras que pueda hacer posible su recuperación, categorización y análisis automático.

Este crecimiento acelerado de grandes cantidades de datos clínicos no estructurados se debe a la adopción generalizada del "Expediente Clínico Electrónico (ECE)". El aprovechamiento adecuado de esta información tiene el potencial de llevar a cabo la atención clínica asistida.

La oncología es una especialidad clínica que genera grandes cantidades de notas o reportes médicos en texto no estructurado. Esta información, en muchos casos, no es analizada en su totalidad, ni procesada para apoyar la toma de decisiones, como un diagnóstico temprano de cáncer, tratamiento oportuno o monitoreo constante de pacientes oncológicos diagnosticados.

En este dominio, el paciente se convierte en un factor clave y punto focal de toda herramienta o sistema de tratamiento de información. En un futuro, los responsables de las decisiones clínicas pueden apoyarse de estas herramientas de análisis de textos clínicos para mejorar los diagnósticos y evitar errores en esta decisión crítica. La idea es mejorar la calidad de vida de los pacientes mediante una adecuada y oportuna decisión clínica, ya sea a pacientes hospitalizados o pacientes ambulatorios. Tanto para los pacientes hospitalizados como para los pacientes ambulatorios, es de vital importancia realizar un diagnóstico sin errores y de manera oportuna, además de monitorear su estado de salud de manera automatizada. El tratamiento de esta información no estructurada involucra un gran reto para el PLN debido a la gran diversidad de estructuras y fenómenos del lenguaje presentes en estos reportes o notas clínicas. Sin embargo, el aprendizaje supervisado apoyado por técnicas de PLN puede hacer posible la clasificación de notas o reportes clínicos para apoyar el diagnóstico de cáncer. Una nota de admisión se genera cuando un paciente ya tiene abierto un expediente clínico y en éste se describe su padecimiento actual.

El diagnóstico y tratamiento temprano del cáncer, así como su monitoreo constante y oportuno, mejora significativamente la calidad de vida de los pacientes hospitalizados y ambulatorios. Herramientas y enfoques computacionales han sido propuestos para analizar, automáticamente, los textos de notas y reportes clínicos. Por ello, a continuación, se presentan diversos trabajos con la finalidad de extraer información, categorizar y clasificar textos clínicos.

La clasificación de textos clínicos ha sido abordada por [Garla, 2012] quienes utilizan una ontología del dominio de la medicina y mediciones de similitud semántica para mejorar la clasificación de textos clínicos de diversas enfermedades como asma, depresión, diabetes, obesidad, hipertensión, entre otras; en [Garla et al., 2013] clasifican textos clínicos con un enfoque semi-supervisado basado en Máquinas de Soporte Vectorial de Laplace, la idea principal es etiquetar reportes ultrasónicos ante la presencia o ausencia de lesiones hepáticas potencialmente malignas y que requieren un seguimiento hospitalario; en [Sarker, 2015] se propone un enfoque para la anotación de

sentencias a partir de reportes clínicos utilizando una gran cantidad de características semánticas y sintácticas; en [Parlak, 2015] se propone una metodología para la clasificación de documentos médicos basada en una lista enfermedades, la relevancia del trabajo recae en que se trata de una clasificación multi-etiqueta, es decir, que un documento médico puede pertenecer o describir más de una enfermedad, utilizando tres algoritmos de clasificación: redes bayesianas, árboles de decisión C4.5 y *Random Forest*. Finalmente, en [Zhao et al., 2013] también consideran la clasificación de textos clínicos libres con multi-etiquetas, donde los autores exploran las relaciones de las palabras de una definición de las enfermedades para conformar su vector de características y así utilizar un clasificador simple de cadenas.

El agrupamiento de textos clínicos consiste en determinar los grupos y los textos que pertenecen a cada grupo sin contar con un conjunto de textos previamente etiquetados. En este rubro, se presentan los siguientes trabajos: en [Paul, 2013] se expone cómo utilizar el conocimiento del dominio médico en el proceso de agrupamiento con el propósito de predecir la probabilidad de enfermedades, los autores combinan el algoritmo de agrupamiento *k-means* y *k-mode* con conocimiento del dominio; en [Ling et al., 2015] se presenta un sistema para la extracción de nombres de medicamentos y síntomas a partir de notas clínicas con la finalidad de agrupar documentos clínicos, los autores utilizan una matriz de factorización para el agrupamiento de notas clínicas.

La extracción de información a partir de textos clínicos es una tarea que sirve como base para diversas tareas de clasificación, agrupamiento o tratamiento de datos clínicos. En este rubro existen enfoques para la extracción de síntomas para diversas enfermedades como la depresión, hipertensión, diabetes ([Riley, 1997], [Divita et al., 2106], [Ma et al., 2017] y [Shao, 2004]); medicamentos o fármacos para el tratamiento de alguna enfermedad o iteraciones medicamentosas ([Peters et al., 2016], [Kuwayama et al., 2016] y [Paul, 2013]); nombres de enfermedades y relaciones entre ellas con la finalidad de encontrar antecedentes familiares heredables en las notas clínicas [Kumar, 2016] y [Mahmood et al., 2016] y o monitoreo de pacientes [Nguyen y Nguyen, 2015] y [Roberts et al., 2008].

Finalmente, la extracción de eventos, ya sea adversos o eventos generados por un medicamento en un paciente [Santiso et al., 2014], [Jindal, 2013] y [Santiso et al., 2016].

Este artículo presenta un enfoque de clasificación supervisada de reportes clínicos, específicamente, notas de admisión en la especialidad de oncología. El proceso completo involucra el procesamiento del texto de la nota de admisión mediante un pre-procesado, extracción de características lingüísticas (etiquetas gramaticales de las palabras) y finalmente, una ponderación de las características que serán la base del algoritmo de aprendizaje supervisado para determinar el tipo de cáncer descrito en la nota de admisión. El resto del artículo se organiza como sigue. La sección 2 expone los métodos de aprendizaje automático y las técnicas de Procesamiento de Lenguaje Natural utilizados para la predicción del tipo de cáncer descrito en una nota de admisión oncológica. Por su parte, la sección 3 exhibe los resultados obtenidos en la tarea de predicción de la categoría de cáncer a partir del análisis de las notas de admisión y la sección 4 presenta la discusión de resultados. Finalmente, las conclusiones de este artículo son expuestas en la sección 5.

2. Métodos

En esta sección se presenta la arquitectura del enfoque propuesto para la clasificación de reportes clínicos, específicamente, notas de admisión, donde se analiza la descripción del padecimiento actual y se clasifica en alguna de cuatro categorías posibles: cáncer de estómago, cáncer pulmonar, cáncer de pecho o cáncer de piel, con la finalidad de apoyar la toma de decisiones clínicas para el diagnóstico temprano de algún tipo de cáncer. El proceso completo de predicción del tipo de cáncer abarca diversos pasos: el pre-procesado de las notas de admisión (lematización, etiquetado POS, eliminación de palabras vacías); la representación y ponderación de las características lingüísticas; y la clasificación de notas de admisión mediante el algoritmo de aprendizaje supervisado llamado Máquinas de Soporte Vectorial (MSV). En la figura 1 se muestra la arquitectura general del enfoque propuesto para la clasificación de notas de admisión.

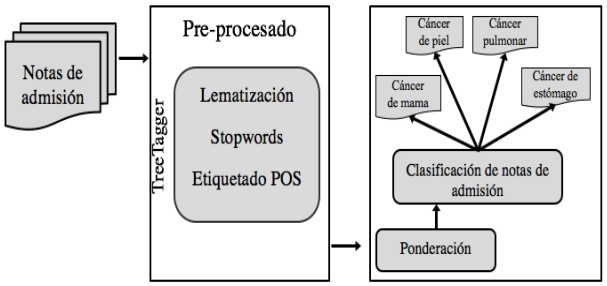

Figura 1 Arquitectura general del enfoque propuesto

Pre-procesado de Notas de Admisión

Las notas de admisión creadas por los médicos presentan una sección de padecimiento actual, donde se describe la evaluación realizada al paciente en cuestión. En ella se describen los signos y síntomas, apariencia y hallazgos del paciente. Este padecimiento actual es descrito como texto no estructurado y puede contener información no relevante. Por ello, se realiza una serie de tareas como pre-procesado con la finalidad de mejorar la calidad de estos textos.

Los textos en la nota de admisión que describen el padecimiento actual del paciente es segmentado, es decir, dividido en palabras (*tokens*), además, se eliminan los caracteres especiales (# $ % & *), puntos, comas y signos (¿, ¡). Las oraciones resultantes son etiquetadas con las categorías gramaticales correspondientes (verbos sustantivos, adjetivos, pronombres y determinantes) mediante la herramienta *TreeTagger* [Helmut, 1995]. Adicionalmente, una lematización se lleva a cabo, la cual consiste en reducir las palabras a sus raíces, eliminando los sufijos, flexiones y conjugaciones de las palabras. La tabla 1 muestra ejemplos de palabras y el resultado que se obtiene de la segmentación, lematización y etiquetado gramatical.

Tabla 1 Lematización y etiquetado de palabras.

Palabra	**Raíz**	**Etiqueta gramatical**
paciente/*patient*	*patient*	NN
tiene/*has*	*have*	VHZ
un/*a*	*a*	DT
severo/*severe*	*severe*	JJ
dolor de cabeza/ *headache*	*headache*	NN
donde NN = Sustantivo; VHZ= verbo "tener" en tercera persona del singular; DT= determinante; JJ = adjetivo.		

Adicionalmente, se lleva a cabo una normalización de las oraciones, mediante la conversión a minúsculas y se eliminan las *stopwords*, palabras que no aportan significado y por lo tanto, no son funcionales para la clasificación del tipo de cáncer. Esta lista de palabras contiene artículos (un, la, los), preposiciones (a, con, de, para) y verbos no funcionales (ser, estar).

Representación y Ponderación de las Características

Se utilizó un conjunto de características para la representación de las notas de admisión. Dos tipos de características en este trabajo: características lingüísticas y uni-gramas de palabras gramaticales.

Las **características lingüísticas** corresponden al número de verbos, sustantivos, conjunciones, determinantes, adjetivos, adverbios, pronombres personales y preposiciones que se encuentran en una nota de admisión. Además, se añaden las siguientes características simples:

a) Longitud promedio de las sentencias en términos de palabras.
b) Presencia de negaciones (no/*no*, negado/*denied*, ni/*neither*, nunca/*never*). Esta es una característica binaria.
c) Presencia de verbos que indican un síntoma (presenta/*present*, tiene/*has*, acude con/*come(s) with*).

Por su parte, los **unigramas** corresponden a la lista de palabras sin repeticiones que contiene una nota de admisión. Se conserva la categoría gramatical por que se ha experimentado con verbos, sustantivos, adjetivos y su combinación.

La ponderación de los **unigramas** se lleva a cabo mediante el modelo de la bolsa de palabras como un vector $V_j = (v_{1j}, v_{2j}, v_{3j})$, el cual consiste en un lexicón o diccionario de las palabras de las notas de admisión. La ponderación consiste determinar los valores de cada palabra de modo que el componente v_{ij} representa la importancia que produce la característica *i*, en la nota de admisión *j* en relación a las palabras de todo el conjunto de notas. La importancia de una palabra (unigrama) está determinado por la fórmula de TF-IDF (Frecuencia de la palabra en la nota de admisión con respecto a la frecuencia de la palabra en todo el

conjunto de notas). Para ello, es necesario obtener, mediante la ecuación 1, el valor de TF (Frecuencia de la palabra), que consiste en el número de veces que una palabra (*t*) aparece en una nota de admisión (*S*).

$$TF(t_i, S_j) = \mathrm{f}(t_i, S_j) \quad (1)$$

Después se obtuvo la frecuencia inversa que determina si el término es común en la colección de notas de admisión clínica y que se obtiene mediante la ecuación 2.

$$IDF(t_i, S_j) = \log \frac{|S|}{1 + |s \in S : t_i \in s|} \quad (2)$$

Esta información se utiliza, entonces, para calcular el valor final de TF-IDF utilizando la ecuación 3.

$$w_{ij} = TF(t_i, S_j) \times IDF(t_i, S_j) \quad (3)$$

Finalmente, una fase de normalización es llevada a cabo a partir de la matriz obtenida de aplicar la ecuación 4.

$$W_{norm} = \frac{w_{ij}}{\sqrt{\sum_{i=0}^{n} |w_{ij}|^2}} \quad (4)$$

Donde *n* representa el número total de notas de admisión y *j* expresa cada nota.

Clasificación de Notas de Admisión

La identificación del tipo de cáncer descrito en la nota de admisión, específicamente, en la sección del padecimiento actual de los pacientes, corresponde a una tarea típica de clasificación de textos. La idea es determinar la categoría de cáncer expresado en la nota. Cuatro tipos de cáncer son considerados en las notas de admisión: cáncer de estómago, cáncer pulmonar, cáncer de pecho y cáncer de piel.

La clasificación de los textos que corresponden al padecimiento actual se basa en el vector ponderado y normalizado de las palabras de cada una de las notas de admisión. Estos vectores son la entrada para el algoritmo de clasificación supervisada Máquinas de Soporte Vectorial, cuyo objetivo es predecir la categoría de la nota basándose en un conjunto de notas de entrenamiento previamente etiquetadas.

La tarea de clasificación de notas de admisión se lleva a cabo mediante el algoritmo de Máquinas de Soporte Vectorial (MSV) [Chang, 2001], el cual ha sido

ampliamente utilizado en la clasificación de textos con etiquetas simples. Este clasificador construye un conjunto de hiperplanos en un espacio n-dimensional con los textos de las notas de entrenamiento, estos hiperplanos son utilizados para predecir la clase de las nuevas notas de admisión.

La idea es evaluar la tarea de clasificación, combinando las diversas características (lingüísticas y n-gramas) con el algoritmo MSV y la ponderación de las palabras (TF-IDF), para encontrar la mejor configuración en cuanto a precisión y cobertura. La implementación del algoritmo de clasificación se ha llevado a cabo mediante la herramienta WEKA [Garner, 1995].

3. Resultados

En este artículo realizamos una experimentación de la tarea de clasificación de reportes clínicos (notas de admisión) para la detección temprana de cáncer con la finalidad de apoyar la toma de decisiones con respecto a los diagnósticos. La experimentación se basa en n-gramas de palabras y características lingüísticas de las notas. La evaluación del enfoque de clasificación de notas de admisión se lleva a cabo con la base de datos denominada MIMIC-II [Saeed, 2011]. La base de datos de MIMIC-II contiene, entre otros datos, notas clínicas con texto no estructurado en inglés de aproximadamente 40 000 estancias en Unidades de Cuidados Intensivos (UCI) de casi 33 000 pacientes en el Centro Médico Beth Israel Deaconess (BIDMC) en Boston, Massachusetts, entre 2001 y 2008.

Para cada hospitalización, seleccionamos todas las notas de la UCI, que describen la nota de admisión del médico. Estas forman lo que llamamos un conjunto de notas de admisión. Estas notas corresponden a pacientes, los cuales son identificados través del campo HADM_ID de las hospitalizaciones durante las cuales se asignó o realizó un diagnóstico de interés. Este diagnóstico es de suma importancia para nuestro enfoque ya que proporciona una sola etiqueta para cada nota de admisión, esto se soluciona con el clasificador Máquinas de Soporte Vectorial (MSV) con etiquetas simples.

A partir del conjunto total de notas de admisión se extraen solo las que pertenecen a cuatro categorías (cáncer de estómago, cáncer pulmonar, cáncer de pecho o

cáncer de piel), debido a que presentan una mayor cantidad de notas. Un total de 1078 notas de admisión son extraídas, distribuidas de la forma como se muestran en la tabla 2, la cual expone la cantidad de notas para cada categoría de diagnóstico. Después, estas notas serán divididas en dos conjuntos: entrenamiento y pruebas.

Tabla 2 Distribución de notas de admisión por diagnóstico.

Diagnóstico	Cantidad de notas de admisión
cáncer de estómago	220
cáncer pulmonar	405
cáncer de pecho	223
cáncer de piel	230
Total	**1078**

Los vectores de características son extraídos a partir de este conjunto de 1078 notas de admisión. Luego, estos vectores son divididos en un conjunto de entrenamiento del clasificador y otro conjunto de prueba para validar la eficiencia de la tarea de clasificación. El conjunto de entrenamiento corresponde con el 70% de las notas para tener un total de 754, mientras que el resto corresponde al conjunto de prueba, el cual está constituido por 324 notas de admisión.

La experimentación consiste en utilizar el algoritmo de clasificación denominado Máquinas de Soporte Vectorial con la ponderación TF-IDF. Todos los experimentos se llevaron a cabo con los siguientes parámetros: parámetro de complejidad (número de hiperplanos a construir): -C 1; parámetro gama (tipo de kernel a utilizar): -K PolyKernel; tamaño de la memoria cache a utilizar: -C 250007; parámetro de tolerancia: -L 0.001.

Para la etapa de evaluación del clasificador, se provee un conjunto de pruebas, mutuamente excluyente del conjunto de entrenamiento, que consiste en 324 reportes clínicos (notas de admisión). Se utilizan dos métricas prácticas para el análisis de la clasificación de notas de admisión: porcentaje de instancias clasificadas correctamente (C) para cada categoría y el porcentaje de instancias clasificadas incorrectamente (I). Todos los experimentos se realizaron sobre el conjunto etiquetado de manera única en cuatro tipos de diagnósticos de cáncer: estómago, pulmonar, pecho, piel.

Una experimentación exhaustiva y comparativa entre las características lingüísticas y los unigramas de palabras por categoría gramatical es desempeñada. Las categorías gramaticales a evaluar son sustantivos, verbos y adjetivos. La tabla 3 muestra los resultados de la clasificación utilizando características lingüísticas por cada categoría, las cuales fueron descritas en la sección 2.

Tabla 3 Resumen de clasificación de notas médicas utilizando características lingüísticas.

Tipos de diagnóstico	% Correctas	% Incorrectas
Cáncer de estómago	45.4	54.6
Cáncer pulmonar	59.1	40.9
Cáncer de pecho	38.1	61.9
Cáncer de piel	47.8	52.2
Promedio	47.6	52.4

La tabla 4 muestra los resultados de la clasificación utilizando el léxico de unigramas de verbos, sustantivos y adjetivos como características para la clasificación de las notas clínicas basada en el diagnóstico de tipo de cáncer.

Tabla 4 Resumen de clasificación de notas médicas con características gramaticales.

Característica gramatical	Sustantivo		Verbo		Adjetivo	
	% C	% I	% C	% I	% C	% I
Cáncer de estómago	68.3	31.7	56.3	43.7	64.5	35.5
Cáncer pulmonar	76.2	23.8	62.1	37.9	71.7	28.3
Cáncer de pecho	73.1	26.9	52.4	47.6	61.7	38.3
Cáncer de piel	71.0	29.0	50.4	49.6	63.9	36.1
Promedio	72.1	27.9	55.3	44.7	65.4	34.6

4. Discusión

Los resultados presentados en las tablas 3 y 4 muestran, en resumen, que resulta mejor la clasificación de notas de admisión utilizando características gramaticales que utilizar características lingüísticas simples como la longitud de las oraciones, la presencia de negaciones y la presencia de verbos que indican síntomas. Además, la tabla 4 demuestra que a partir de las categorías gramaticales, se obtienen los mejores resultados utilizando sustantivos, con los cuales se logra un porcentaje promedio de 72.1% de instancias correctamente

clasificadas para las cuatro categorías. Por lo tanto, se comprueba que este comportamiento se debe a que los sustantivos describen nombres de enfermedades, nombres de medicamento, nombres de síntomas y cualquier entidad nombrada dentro de las notas de admisión. Esto quiere decir, que al detectar cualquier entidad nombrada como un sustantivo, el algoritmo divide con mejor precisión las cuatro categorías de diagnósticos de cáncer.

5. Conclusiones

En este artículo se ha presentado un enfoque para la clasificación de notas cínicas utilizando el algoritmo de aprendizaje automático denominado Máquinas de Soporte Vectorial (MSV) para predecir la categoría o tipo de cáncer descrito en el padecimiento actual de cada una de las notas de admisión.

El procesamiento de las notas de admisión incluye diversas tareas. Primero, un pre-procesado de las notas clínicas que involucra una lematización, eliminación de palabras vacías y etiquetado gramatical de partes de la oración. Además, las notas de admisión son caracterizadas como un conjunto de características lingüísticas y gramaticales, además de utilizar la medida TF-IDF para la ponderación de los términos o características. Finalmente, una clasificación de las notas de admisión mediante el algoritmo MSV es desempeñada sobre el conjunto de notas de admisión.

Las principales aportaciones de este trabajo son a) el enfoque de clasificación de notas de admisión mediante el algoritmo Maquinas de Soporte Vectorial; b) la caracterización de las notas mediante la longitud de las oraciones, la presencia de palabras de negación y la presencia de verbos que indican síntomas, además de presentar, como otra alternativa de caracterización, los lexicones agrupados por categorías gramaticales verbos, sustantivos y adjetivos; c) el descubrimiento de la mejor alternativa de clasificación de notas clínicas para apoyo en el diagnóstico de cáncer, mostrando que los sustantivos caracterizan mejor a los grupos de cáncer elegidos, debido a su propiedad para expresar nombres de enfermedades, nombres de medicamento y nombres de síntomas. El enfoque obtenido puede ser de gran utilidad al generar una herramienta para el apoyo de diagnósticos de

cáncer y es posible extenderlo a diversas enfermedades como diabetes, hipertensión, entre otras.

Como trabajo futuro, se puede experimentar con diversas características del lenguaje utilizado en las notas clínicas como la formación de frases, ya que un nombre de un medicamento, enfermedades o síntoma puede estar expresado como un sustantivo compuesto, lo que se le conoce como sintagma o frase nominal. Además, experimentar con listas de síntomas para cada categoría de cáncer podría ser de gran utilidad para la clasificación de notas clínicas.

6. Bibliografía y referencias

[1] Chang, Ch., Lin, Ch. LIBSVM, A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, pp. 27-28, 2001.

[2] Divita, G., Carter, M. E., Tran, L. T., Redd, D., Zeng, Q. T., Duvall, S., & Gundlapalli, A. V. v3NLP Framework: Tools to Build Applications for Extracting Concepts from Clinical Text. eGEMs, vol. 4, no. 3, 2016.

[3] Garla, V. N., & Brandt, C., Ontology-guided feature engineering for clinical text classification. Journal of biomedical informatics, vol. 45, no. 5, pp. 992-998, 2012.

[4] Kumar, L. S., & Padmapriya, A. Evidence based subsequent disease extraction from EMR Health Record by Grade Measure. In IEEE Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1-5, 2016.

[5] Garla, V., Taylor, C., & Brandt, C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. Journal of biomedical informatics, vol. 46, no. 5, pp. 869-875, 2013.

[6] Garner, S.R. Weka: The Waikato environment for knowledge analysis. In: Proc. of the New Zealand Computer Science Research Students Conference, pp. 57-64, 1995.

[7] Helmut, S., Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland, 1995.

[8] Jindal, P., & Roth, D., Extraction of events and temporal expressions from clinical narratives. Journal of biomedical informatics, vol. 46, pp. 13-19, 2013.

[9] Kuwayama, K., Miyaguchi, H., Iwata, Y. T., Kanamori, T., Tsujikawa, K., Yamamuro, T., & Inoue, H. Three-step drug extraction from a single sub-millimeter segment of hair and nail to determine the exact day of drug intake. Analytica Chimica Acta, 948, pp. 40-47, 2016.

[10] Ling, Y., Pan, X., Li, G., & Hu, X., Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. IEEE transactions on nanobioscience, vol. 14, no. 5, pp. 500-504, 2015.

[11] Ma, L., Wang, Z., & Zhang, Y., Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining. In International Symposium on Bioinformatics Research and Applications, pp. 325-330, 2017.

[12] Mahmood, A. A., Wu, T. J., Mazumder, R., & Vijay-Shanker, K., Dimex: A text mining system for mutation-disease association extraction. PloS one, vol. 11, no. 4, 2016.

[13] Nguyen, M. T., & Nguyen, T. T., DESRM: a disease extraction system for real-time monitoring. International Journal of Computational Vision and Robotics, vol. 5, no. 3, pp. 282-301, 2015.

[14] Parlak, B., & Uysal, A. K., Classification of medical documents according to diseases. In 23th IEEE Signal Processing and Communications Applications Conference (SIU), pp. 1635-1638, 2015.

[15] Paul, R., & Hoque, A. S. M. L., Clustering medical data to predict the likelihood of diseases. In IEEE Fifth International Conference on Digital Information Management (ICDIM), pp. 44-49, 2013.

[16] Paul, M. J., & Dredze, M., Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. In HLT-NAACL, pp. 168-178, 2013.

[17] Peters, S. A., Jones, C. R., Ungell, A. L., & Hatley, O. J., Predicting drug extraction in the human gut wall: assessing contributions from drug

metabolizing enzymes and transporter proteins using preclinical models. Clinical pharmacokinetics, vol. 55, no. 6, pp. 673-696. 2016.

[18] Riley, D. S., Extracting symptoms from homoeopathic drug provings. British Homoeopathic Journal, vol. 86, no. 4, pp. 225-228. 1997.

[19] Roberts, A., Gaizauskas, R., & Hepple, M., Extracting clinical relationships from patient narratives. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics, pp. 10-18, 2008.

[20] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access ICU database. Crit Care Med, vol. 39, pp. 952–60, 2011.

[21] Santiso, S., Pérez, A., Gojenola, K., Taldea, I. X. A., Casillas, A., & Oronoz, M. Adverse Drug Event prediction combining shallow analysis and machine learning. In Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) EACL, pp. 85-89, 2014.

[22] Santiso, S., Casillas, A., Pérez, A., Oronoz, M., & Gojenola, K. Document-level adverse drug reaction event extraction on electronic health records in Spanish. Procesamiento del Lenguaje Natural, no. 56, pp. 49-56, 2016.

[23] Sarker, A., & Gonzalez, G., Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, 53, 196-207, 2015.

[24] Shao, Y., & Nezu, K., Extracting symptoms of bearing faults in the wavelet domain. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, vol. 218, no. 1, pp. 39-51. 2004.

[25] Zhao, R. W., Li, G. Z., Liu, J. M., & Wang, X. Clinical multi-label free text classification by exploiting disease label relation. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 311-315, 2013.