Sentiment Analysis on Tweets related to infectious diseases in South America

Short Paper

José Antonio García-Díaz
Departamento de Informática y Sistemas
Universidad de Murcia
Murcia, Spain
joseantonio.garcia8@um.es

Óscar Apolinario-Arzube
Facultad de Ciencias Matemáticas y Físicas
Universidad de Guayaquil
Guayaquil, Ecuador
oscar.apolinarioa@ug.edu.ec

José Medina-Moreira
Facultad de Ciencias Matemáticas y Físicas
Universidad de Guayaquil
Guayaquil, Ecuador
jose.medinamo@ug.edu.ec

Harry Luna-Aveiga
Facultad de Ciencias Matemáticas y Físicas
Universidad de Guayaquil
Guayaquil, Ecuador
harry.lunaa@ug.edu.ec

Katty Lagos-Ortiz
Facultad de Ciencias Agrarias
Universidad Agraria del Ecuador
Guayaquil, Ecuador
klagos@uagraria.edu.ec

Rafael Valencia-García
Departamento de Informática y Sistemas
Universidad de Murcia
Murcia, Spain
valencia@um.es

ABSTRACT

Infectious diseases have a huge social and economic impact. They are caused by pathogenic microorganisms such as bacteria, viruses, parasites or fungi and they can be transmitted, directly or indirectly, from one person to another or from animals to humans (Zoonoses). Nowadays it is very important to detect the infectious diseases as soon as possible to prevent critical problems for the society. In this work we propose an approach for the sentiment classification of tweets related to infectious diseases. This kind of systems could help health professionals to know how society respond to advances in the treatment of these diseases. In addition, a comparison was made of the performance of three classification algorithms (J48, BayesNet, and SMO). The results showed that SMO provides better results than BayesNet and J48 algorithms, obtaining an F-measure of 84.4%.

CCS CONCEPTS

• **Information systems** → Sentiment analysis • **Applied computing** → Health care information systems

KEYWORDS

sentiment analysis, infectious diseases, natural language processing

EATIS '18, November 12–15, 2018, Fortaleza, Brazil © 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6572-7/18/11..$15.00
https://doi.org/10.1145/3293614.3293647

ACM Reference format:
J. A. García-Díaz, O. Apolinario-Arzube, J. Medina-Moreira, H. Luna-Aveiga, K. Lagos-Ortiz, and R. Valencia-García. 2018. Sentiment Analysis on tweets related to infectious diseases in South America. In Euro American Conference on Telematics and Information Systems (EATIS '18), November 12–15, 2018, Fortaleza, Brazil. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3293614.3293647

1 INTRODUCCIÓN

Las enfermedades infecciosas son aquellas enfermedades provocadas por microorganismos patógenos que se transmiten i) a través del intercambio de fluidos, ii) a través del medioambiente o iii) a través de agentes portadores como es el caso de la malaria, dengué, zika y chikungunya. Cuando el número de afectados en una determinada zona es superior al esperado durante un periodo de tiempo concreto hablamos de epidemias o brotes. La frecuencia de estos brotes está aumentando en el tiempo debido principalmente i) al crecimiento continuado de la población, ii) a cambios medioambientales, iii) a movimientos migratorios de la población, iv) al turismo, comercio y viajes internacionales, v) a cambios en el comportamiento sexual de los individuos y vi) el uso indiscriminado de insecticidas y compuestos antimicrobianos [1].

La monitorización de enfermedades infecciosas es una actividad que consiste en la observación sistemática de distintos parámetros médicos relacionados con dicha enfermedad a través de la recolección de datos fiables, de calidad, precisos y a tiempo con el fin de desarrollar planes de acción eficaces que sirvan para evitar y contrarrestar sus devastadores efectos, tanto para la economía como para la sociedad [1]. Diversos organismos, como el IHR (International Health Regulations), han indicado la necesidad de crear y fortalecer sistemas de monitorización que permitan identificar eficazmente brotes de enfermedades

contagiosas. La lucha contra este tipo de enfermedades es un problema actual; actualmente se identificaron 37 virus que tendrían la posibilidad de transmitirse entre seres humanos [2]. Además, al analizar la mayoría de los casos recientes, se ha comprobado que la respuesta gubernamental ha sido lenta y descoordinada, siendo desproporcionada e impulsiva en ciertos casos. Estas actuaciones han desacreditado en gran medida a las instituciones en su conjunto [3].

Con el objetivo de poder facilitar a los profesionales de la salud la información necesaria para poder tomar decisiones efectivas, en este trabajo se presenta un estudio de análisis de sentimientos en español de tuits relacionados con los virus zika, malaria y chikungunya en Sudamérica. Con este análisis se pretende determinar la polaridad de textos relacionados con estas enfermedades y así poder determinar la percepción social ante sucesos relacionados con las mismas en entornos sociales.

2 ESTADO DEL ARTE

2.1 Análisis de Sentimientos

El Análisis de Sentimientos es una rama del Procesamiento del Lenguaje Natural (PLN) que se basa en inferir la polaridad subjetiva que tendrían los individuos ante un texto específico con respecto a un dominio concreto. El primer estudio respecto al Análisis de Sentimientos consistió en el análisis de la opinión generalizada de películas de cine a través de un repositorio de comentarios de usuarios [4].

El Análisis de Sentimientos se suele abordar con técnicas de aprendizaje supervisado, es decir, a través del estudio de un conjunto de datos previamente clasificado. Esta clasificación previa se puede hacer de forma automática o manual [5]. Las clasificaciones automáticas se pueden realizar cuando los textos a clasificar disponen de algún tipo de metadato, como puede ser un ranking numérico o la presencia de emoticonos entre otras estrategias. En cambio, el proceso de clasificación manual, aunque más lento, tiende a ser más efectivo ya que los humanos somos capaces de abarcar una clasificación más compleja.

La polaridad expresada hacia un dominio concreto puede registrarse de distintas maneras. La manera más habitual de hacerlo es mediante una clasificación binaria: positiva o negativa. Esta clasificación más polarizada tiende a funcionar mejor que otras clasificaciones menos polarizadas, como es el sistema de valoración de 5 rangos: i) muy negativo, ii) negativo, iii) neutro, iv) positivo y v) muy positivo, puesto que recoge con mayor consenso la opinión generalizada de los usuarios.

El Análisis de Sentimientos se puede efectuar a i) nivel de documento, ii) nivel de frase o iii) nivel de aspectos [6]. En una clasificación a nivel de documento, la polaridad se extrae de manera global a todo el texto. En una clasificación a nivel de frase, los documentos son divididos en frases que son analizadas de manera individual. Por último, la clasificación a nivel de aspectos intenta realizar un análisis más minucioso tratando de identificar previamente distintos conceptos expresados en el documento. Sin embargo, independientemente del nivel de profundidad al que se quiera llegar, la tarea de clasificar el sentimiento de textos es complicada [7] debido a que i) distintas personas podrían clasificar los textos de manera distinta, por lo que para intentar luchar contra la subjetividad de las opiniones es necesario disponer de un catálogo suficientemente amplio de opiniones y ii) los sentimientos son muy dependientes del dominio, por lo que es difícil generalizar este tipo de soluciones para distintas situaciones.

A nivel técnico, el Análisis de Sentimientos puede ser efectuado utilizando diversas estrategias. Un primer enfoque consiste en calcular la aparición de ciertas palabras en el texto. Este enfoque es conocido como Orientación Semántica. Para ello, se suelen emplear repositorios de palabras relacionadas con sentimientos tales como SentiWord Net [8]. Por otro lado, se puede realizar un enfoque basado en técnicas de machine-learning. Según este método, los sentimientos son clasificados empleando técnicas que construyen un clasificador a partir de un corpus previamente etiquetado y través del estudio de ciertas características extraídas de los mismos, como pueden ser un modelo de bolsa de palabras o a través de características lingüísticas. El enfoque de machine-learning suele dar muy buen resultado y ha sido probado con éxito en distintos trabajos (por ejemplo [9]). Sin embargo, las técnicas basadas en machine-learning dependen en gran medida de la calidad del corpus que se emplee como entrenamiento; conseguir un campus entrenado suele ser una tarea compleja ya que en muchos casos es necesario un entrenamiento manual por parte de expertos. Además, es muy importante conseguir que el corpus sea representativo del dominio que se trata de analizar.

La gran mayoría de estudios realizados en el campo del Análisis de Sentimientos se han realizado sobre el lenguaje inglés, aunque cada vez se pueden encontrar en la literatura más trabajos que amplían este estudio a otros lenguajes como el español (por ejemplo [10] y [11]). El Análisis de Sentimientos requiere de un mayor esfuerzo en casos de textos que combinan diversos lenguajes.

2.2 Detección temprana de enfermedades infecciosas

La detección temprana de epidemias de enfermedades infecciosas es llevada a cabo por una red de profesionales de la sanidad pública orientados a la vigilancia e investigación de epidemias. Esta red, conocida como red de médicos centinela, suele realizar estudios periódicos con objeto de encontrar mutaciones de virus comunes o nuevos casos de enfermedades infecciosas. Esta detección tiene lugar, normalmente, una vez que los médicos atienden a los pacientes de hospitales o al hacer controles rutinarios en puestos de gran afluencia de tráfico como aduanas o puestos de fronteras. La automatización de estos procesos de detección temprana se conoce como Infoveillance, que hace referencia al uso continuo y sistemático de información para la evaluación de la salud pública [12]. Este enfoque ha sido usado principalmente con el análisis de textos de redes sociales. En [13], los autores llevaron a cabo un análisis de textos en Twitter en busca de contraindicaciones y riesgos en ciertos productos del mercado, como pueden ser productos cosméticos o productos alimentarios. En [14], los autores realizaron un estudio acerca de medir la eficiencia del uso de información en redes sociales con

el objetivo de mejorar la calidad de la detección temprana de enfermedades analizando tópicos relacionados con la salud de dominio público. La información fue estudiada también a posteriori presentado especial atención a comparar textos ocasionados en el lugar de origen con mensajes en áreas circundantes. Por otra parte, un estudio acerca de la percepción social relacionado con los riesgos de la salud fue llevado a cabo en [15]. Este estudio sirvió para generar un posible catálogo de buenas prácticas a la hora de avisar a la población en caso de tener que informar acerca de este tipo de desastres. El uso de Twitter para estudiar temas relacionados con el virus del zika ya ha sido previamente aplicado en [16] en donde los autores realizaron un análisis de textos para determinar el volumen de datos estudiados relacionados con los síntomas a corto y largo plazo.

3 EXPERIMENTO

El experimento descrito en el presente trabajo está relacionado con la creación y validación de un mecanismo clasificador del análisis de sentimiento de un corpus de tuits relacionados con los virus del zika, dengué y chikungunya. El objetivo es mejorar la calidad de la información para que profesionales sanitarios puedan identificar, de forma anticipada, posibles casos de enfermedades infecciosas, así como su evolución y contagio. Con el desarrollo de este tipo de sistemas de análisis y minería de textos se pueden inferir pautas de comportamiento y tendencias que indiquen los casos que deban de ser alertados a los profesionales de la salud pública. Además de la creación de un repositorio de datos dónde se puedan aplicar técnicas de aprendizaje a posteriori, y así poder adaptar el modelo de predicción conforme a nuevos datos.

3.2 Descripción del experimento

En este trabajo hemos aplicado un análisis de sentimientos al dominio de enfermedades infecciosas dentro de un ámbito geográfico concreto. Para cumplir con este objetivo, se han diseñado distintas herramientas. En primer lugar, una aplicación que nos permite generar corpus a través la red social Twitter. Esta herramienta, llamada UMUCorpusClassifier se encarga de obtener distintos tuits a partir de las posibilidades que ofrece la API de Twitter como son: i) la búsqueda por palabras clave, ii) hashtags, iii) área geográfica, o iv) fecha entre otros. Esta misma herramienta, permite además que un grupo de usuarios puede realizar una clasificación manual de los corpus. En la Fig 1 se muestra una captura de pantalla de esta aplicación.

Figura 1. Captura de pantalla de la aplicación UMUCorpusClassifier en vista desde un dispositivo móvil

Merece la pena destacar que el contenido obtenido por Twitter suele presentar ciertos inconvenientes: i) es difícil de cuantificar y medir su calidad ya que, este tipo de métricas dependen mucho del contexto, ii) suele ser bastante heterogéneo, ya que muchas aplicaciones proveen datos de diferentes maneras como *tags* o *reviews* y iii) puede estar sujeto, como otras fuentes de información, a problemas de calidad, objetividad y privacidad. Sin embargo, y pesar de estos inconvenientes, Twitter es ampliamente utilizado en este tipo de estudios debido al gran volumen de datos que tiene, a su disponibilidad pública y a que ofrece una fuente de información con datos actualizados con las últimas tendencias.

Para este estudio, se obtuvo un corpus acerca de enfermedades infecciosas relacionadas con el zika a partir de tuits publicados durante el periodo de cuatro semanas procedentes principalmente de América Latina. En concreto, la cadena de búsqueda que usamos para obtener los tuits fue la siguiente: *#zica OR #zika OR #dengue OR #chikungunya OR "zica" OR "zika" OR "dengue" OR "chikungunya"* y el radio geográfico donde obtuvimos los tweets fueron las coordenadas de latitud -0.1596997 y longitud -78.452125313 con un radio de 1.500 km que englobaba los países de Ecuador, Colombia, Costa Rica, Panamá y la parte occidental de Venezuela. Se aplicaron diversos filtros a los tuits obtenidos con objeto de obtener el corpus más representativo posible. Estos filtros automáticos detectaron tuits repetidos y tuits con información no relevante, como aquellos que contienen únicamente URLs o menciones.

El proceso de clasificación manual fue llevado a cabo por un grupo de 20 alumnos. Previo al proceso de clasificación, este grupo de alumnos contó con i) una sesión de formación, ii) un manual de usuario de la aplicación y iii) un documento con ejemplos ilustrativos de criterios de clasificación. El proceso de clasificación manual fue monitorizado sistemáticamente a partir de una revisión manual cada semana. Una vez analizadas las últimas clasificaciones, se suministra feedback a los alumnos en función de las conclusiones obtenidas. Como ejemplo de este feedback, se indicó a los alumnos que debían de hacer caso omiso

a enlaces que podían acompañar a los tuits, debiéndose fijar únicamente en el texto del tuit obviando el contenido accesible a través de hipervínculos, como ocurría con el tweet "Me gustó un video de @YouTube https://t.co/d1BhXPA... ". Al final de este periodo, los alumnos cumplieron el objetivo inicialmente planteado de clasificar manualmente un total de 80.000 tweets en muy negativo, negativo, neutro, positivo, muy positivo y no aplica. Es importante destacar que distintos usuarios clasificaron de forma independiente los mismos tuits. Podemos ver una gráfica de la evolución de este proceso en la Fig. 2.

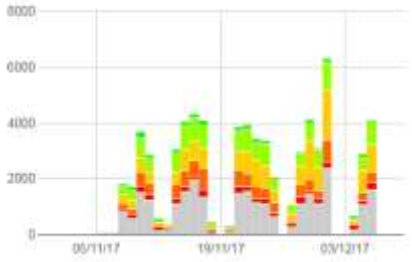

Figura 2. Clasificación de los tweets durante el tiempo

Como podemos observar en la figura anterior, el flujo de trabajo fue constante y continuado, pudiéndose apreciar, sin embargo, un menor rendimiento durante los fines de semana. Como nota adicional, añadir que las estadísticas representadas en este gráfico están obtenidas con horario español, por lo que se observar pequeños picos de trabajo durante fines de semana.

Una vez el corpus fue obtenido, se procedió con la asignación de una polaridad concreta a cada tuit. Para calcular dicha polaridad, por cada tuit se añadía un punto positivo cuando tenía una valoración positiva o muy positiva, o bien se restaba un punto cuando se encontraba una valoración negativa o muy negativa. Esta puntuación nos sirvió para descartar todos los tuits que obtuvieron un valor cercano a cero puesto que esto nos indicaba que había sido clasificado de manera contradictoria por distintos usuarios. Una vez obtenidos y filtrados los tuits se compiló un corpus con un subconjunto balanceado de 500 tweets clasificados manualmente como positivos y otros 500 clasificados como negativos para el experimento.

Este corpus fue usado como entrada de la herramienta UMUTextStats, explicada detalladamente en la siguiente sección, para obtener distintas características léxicas del texto adaptadas al idioma español. Una vez obtenidos estos datos se utilizó el marco de trabajo WEKA [17] aplicando distintos algoritmos. En concreto se usaron los algoritmos SMO, J48 y BayesNet. WEKA fue configurado para realizar una validación cruzada de 10 veces (*10-fold cross validation*) y así poder valorar la precisión del modelo predictivo. Una validación cruzada consiste en partir el corpus de entrenamiento en dos subgrupos, un subgrupo que se utilizará para entrenar y otro conjunto que se usará para comprobar la calidad del modelo aprendido.

3.3 UMUTextStats

La aplicación UMUTextStats permite el estudio de características léxicas, morfológicas, semánticas, emocionales y cognitivas de textos escritos en lenguaje español. La salida de esta aplicación consiste en un vector con las diferentes características de cada texto. Esta aplicación utiliza analizadores léxico-morfológicos y sintácticos junto con un conjunto de diccionarios para procesar los textos. En el momento de este análisis, la aplicación contaba con un total de 112 diccionarios. Estos diccionarios suman un total aproximado de 50.000 palabras y de expresiones regulares agrupadas en 125 características agrupadas en i) categorías léxico-morfológicas y ii) categorías semánticas. Las características léxico-morfológicas incluyen porcentajes acerca del número de palabras, artículos, sentencias, verbos auxiliares, signos de puntuación, etc. Por otro lado, las características semánticas incluyen grupos de palabras relacionadas con emociones tanto positivas como negativas, así como otras actividades sociales. Estos diccionarios se han conformado desde distintas fuentes de datos tales como diccionarios públicos, tesauros y ontologías. Además, para este experimento, se han definido una serie de dimensiones relacionadas con Twitter, con la idea de obtener información acerca del total de hashtags, emoticonos, menciones, etc.

Esta aplicación utiliza un enfoque similar a LIWC (*Language Inquiry and Word Count*) [18] que es una aplicación ampliamente utilizada que clasifica las palabras de un texto en categorías del dominio psicolingüístico. Sin embargo, debido a la necesidad de definir nuevas características específicas para el idioma español, se desarrolló la herramienta de UMUTextStats. Además, pese a los excelentes resultados obtenidos por LIWC en múltiples dominios, consideramos que esta herramienta está enfocada únicamente a características del idioma inglés y, aunque cuanta con diccionarios publicados en español, creemos que no es capaz de medir con precisión algunas características del español. Por ejemplo, la aparición del símbolo de abrir interrogación "¿" puede usarse para indicar que textos escritos en español utilizan un enfoque formal, ya que este símbolo no se suele usar en lenguaje informal, especialmente en sitios donde existe un límite de caracteres máximo a escribir. Además, las oraciones en inglés suelen ser mucho más cortas y escuetas. En cambio, en países de habla hispana, las oraciones suelen ser mucho más largas por lo que decidimos también añadir dimensiones que miden el número total de sentencias que aparecen en los textos. En la Tabla 1 se muestran algunas de las características de UMUTextStats.

Tabla 1. Características de UMUTextStats

Conjunto	Categorías	Ejemplo
Léxico morfológico	Pronombres	Yo, nosotros, tu
	Artículos	Él, la, los, las
	Signos de puntuación	Puntos, comas, signos de porcentaje
Semántico	Emociones positivas	Feliz, bonito, bueno
	Emociones negativas	Odio, enemigo, feo

Twitter	Hashtags Menciones	#zika @cuenta_de_twitter

3.3 Resultados del experimento

Los resultados obtenidos por el experimento tuvieron un éxito, en su mejor resultado, de una medida F de 84.4% usando el algoritmo clasificador SMO con una precisión y una exhaustividad del mismo valor. El resultado de todos los experimentos puede verse en la Tabla 2 que se muestra a continuación.

Tabla 2. Resultados del experimento

Algoritmo	*Precisión*	*Sensibilidad*	*Medida-F*
SMO	**0.844**	**0.844**	**0.844**
BayesNET	0.785	0.776	0.774
J48	0.813	0.813	0.813

Como podemos observar en la Tabla 2, los resultados obtenidos que ofrecen mejor resultado son los obtenidos con el clasificador SMO con un total de 84.4% de medida F. El peor clasificador fue BayesNET con un valor de 77.4%. Es importante observar que los valores obtenidos en los clasificadores de SMO y J48 tiene un valor de precisión muy estable, por lo que éste método es bastante fiable.

4 CONCLUSIONES Y TRABAJO FUTURO

En vista de los resultados de este experimento, podemos concluir que aplicar técnicas de análisis de sentimientos al dominio de enfermedades infecciosas puede ser efectivo. Mediante la obtención de grandes volúmenes online sería posible generar alertas y notificaciones para que las autoridades sanitarias fueran capaces de efectuar una pronta detección de estas y así poder mitigar sus efectos.

Gracias a esta información estratégica, las administraciones públicas son capaces de diseñar sistemas de alertas eficaces que mitiguen los efectos concernientes a la salud pública. Además, la información recolectada sirve para documentar las intervenciones de las administraciones públicas con respecto a estas enfermedades y así retroalimentar este proceso con objeto de mejorarlo sucesivamente.

Los resultados obtenidos en el experimento son prometedores, pero el conjunto de tuits empleados en el experimento es bastante reducido. Está planificado como trabajo futuro realizar un experimento con muchos más tuits para poder afianzar estos resultados con un conjunto más amplio de muestras. Por otro lado, se pretende ampliar este sistema con un módulo de detección de aspectos relacionados con las enfermedades infecciosas como las enfermedades, síntomas y tratamientos para que, además de poder detectar el sentimiento podamos detectar posibles nuevos síntomas de las enfermedades. Por último, se pretende mejorar los diccionarios de UMUTextStats y adaptarlos a otros idiomas como el portugués.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por la Agencia Española de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto KBS4FIA (TIN2016-76323-R).

REFERENCIAS

[1] V. Carchiolo, A. Longheu and M. Malgeri, "Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics". In International Conference on Information Technology in Bio-and Medical Informatics, 2015, pp. 16-24.

[2] M. E. Woolhouse, D. T. Haydon, and R. Antia, "Emerging pathogens: the epidemiology and evolution of species jumps". Trends in ecology & evolution, vol. 20(5), pp. 238-244. 2005

[3] G. Alleyne, M. Claeson, D. B. Evans, P. Jha, A. Mills, and P. Musgrove, "Disease control priorities in developing countries", World Bank/OUP, 2006

[4] B. Pang, and L. Lee. "Opinion mining and sentiment analysis". Foundations and Trends® in Information Retrieval, vol 2(1–2), pp. 1-135, 2008

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann, 2005.

[6] R. Feldman, "Techniques and applications for sentiment analysis". In Communications of the ACM, Vol 56(4), pp. 82-89, 2005

[7] S. Baccianella, A. Esuli and F. Sebastiani: "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In: LREC. pp. 2200–2204, 2010.

[8] R. Hsu, B. See, A. Wu,:" Machine learning for sentiment analysis on the experience project", 2010.

[9] M. del P. Salas-Zárate, M. A. Paredes-Valverde, M. A. Rodriguez-García, R. Valencia-García and G. Alor-Hernández: "Automatic detection of satire in Twitter: A psycholinguistic-based approach". Knowledge-Based Syst. Vol. 128, pp. 20–33, 2017.

[10] M. del P. Salas-Zarate, M. A. Paredes-Valverde, J. Limon, D. A. Tlapa and Y. A. Báez, Y. A. "Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods". J. UCS, vol. 22(5), pp. 691-708, 2016

[11] M. del P. Salas-Zárate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, A. Almela, and G. Alor-Hernández, "A study on LIWC categories for opinion mining in Spanish reviews" in Journal of Information Science, vol 40(6), pp. 749-760, 2014

[12] C. Chew and G. Eysenbach "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak".in PLoS ONE 5(11): e14118. https://doi.org/10.1371/journal.pone.0014118, 2010

[13] H. Isah, P. Trundle and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis". In Computational Intelligence (UKCI), UK Workshop. pp. 1-7. 2014

[14] B. Ofoghi, M. Mann and K. Verspoor. "Towards early discovery of salient health threats: A social media emotion classification technique". In Biocomputing 2016: Proceedings of the Pacific Symposium pp. 504-515., 2016

[15] J. Hao and H. Dai "Social media content and sentiment analysis on consumer security breaches". In Journal of Financial Crime, vol 23(4), pp 855-869, 2016

[16] A. Khatua. "Immediate and long-term effects of 2016 Zika Outbreak: A Twitter-based study". In e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference, pp. 1-6. 2016

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update". In ACM SIGKDD explorations newsletter, vol 11(1), pp. 10-18., 2009

[18] J. W. Pennebaker, M. E. Francis, and R. J. Booth,. "Linguistic inquiry and word count: LIWC 2001". Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.