PA163-2-CLASHISTVI

CLASIFICACIÓN DE HISTORIAS PROVENIENTES DE FAMILIARES DE VÍCTIMAS Y SOBREVIVIENTES DEL CONFLICTO SOCIOPOLÍTICO DE COLOMBIA

Ferney Leonardo Olaya Bello

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2016

PA163-2-CLASHISTVI

CLASIFICACIÓN DE HISTORIAS PROVENIENTES DE FAMILIARES DE VÍCTIMAS Y SOBREVIVIENTES DEL CONFLICTO SOCIOPOLÍTICO DE COLOMBIA

Autor:

Ferney Leonardo Olaya Bello

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director

Rafael Andrés González Rivera

Comité de Evaluación del Trabajo de Grado

Julio Ernesto Carreño Vargas

Carlos Andrés Barreneche Jurado

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~PA163-2-CLASHISTVI/

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
11,2016

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Jorge Luis Sánchez Téllez

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Angela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortíz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Cada una de las metas a cumplir en nuestras vidas requiere de motivaciones que nos impulsen hacia ellas. En mi caso la noticia de la llegada de mi hijo me ha motivado para avanzar como nunca algo lo hubiera podido lograr. Por esta razón mi agradecimiento inmenso a mi pequeño Joaquín, quien aún sin estar en mis brazos ya llena mi vida de felicidad y motivación para seguir adelante.

Mi esposa, mis padres, mis hermanos y mis sobrinos siempre a mi lado, siempre mi apoyo incondicional. A ellos mi agradecimiento por enseñarme que no existe el momento en el que no cuente con ellos para disfrutar de la vida.

Mi director, Rafael A. Gonzalez, quien me brinda la oportunidad de realizar este trabajo bajo su orientación, quien comparte conmigo tantas ideas que aportan a mi formación y mi conocimiento.

Los evaluadores de este trabajo, que han destinado un espacio de sus agendas para orientarme y brindarme sus valiosas opiniones. Wilson López López, Wilson Herney Chavarro Jiménez y Mireya Camacho.

Contenido

ABSTRACT

The present work illustrates how a software prototype was developed for supervised classification of unstructured information associated with texts from stories of relatives of victims and survivors of the socio-political conflict in Colombia. Natural language processing techniques, such as name entity recognition, allow defining in percentage terms the categories to which a text belongs. The Design Science Research and CRISP-DM methodologies are applied together and a user-focused assessment following the Technology Acceptance Model (TAM) is done.

RESUMEN

El presente trabajo ilustra cómo se desarrolló un prototipo de software para clasificación supervisada de información no estructurada asociada a textos de historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia. Técnicas de procesamiento natural del leguaje, como "name entity recognition", permiten definir en cifras porcentuales las categorías a las que pertenece un texto. Son aplicadas en conjunto las metodologías "Design Science Research" y "CRISP-DM" y se realiza una evaluación enfocada en el usuario siguiendo "Technology Acceptance Model (TAM)".

RESUMEN EJECUTIVO

Actualmente en Colombia entidades gubernamentales como la Defensoría del Pueblo o la Unidad para las víctimas realizan tareas de recolección y divulgación de testimonios de familiares de víctimas y sobrevivientes de la violencia sociopolítica del país. La información que se recolecta es creciente y valiosa en su contenido, teniendo en cuenta que las declaraciones son un requisito de acceso al registro único de víctimas [1].

Surge la necesidad de diseñar e implementar artefactos de software que apoyen la clasificación de innumerables textos, con información no estructurada que proviene de las víctimas del conflicto sociopolítico de Colombia, promoviendo así análisis imparciales de narrativas, brindando alternativas para asignación de recursos, impactando positivamente la generación de informes claves para la toma de decisiones o para su divulgación dentro de los ciudadanos, detectando patrones que apoyen reconciliación y otras acciones en el posconflicto, entre otros múltiples beneficios.

En consecuencia, se propone el diseño y desarrollo de un prototipo de clasificación supervisada de historias asociadas a víctimas del conflicto sociopolítico, el cual utilice técnicas de preparación de textos que sean relevantes dentro del contexto.

Para el desarrollo del proyecto se abordan en conjunto las metodologías Ciencia basada en el diseño y CRISP-DM, se define como objetivo general Identificar un modelo supervisado de clasificación de textos para historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia que contribuya con la ejecución análisis imparciales y se realiza evaluación mediante el modelo de aceptación tecnológica (TAM).

La labor de análisis de texto se apoya en un pipeline que se desarrolla utilizando GATE – General architecture for text engineering. En donde se busca alcanzar reconocimiento de entidades a partir Gazetteer definidos por categorías. Cada una de las palabras dentro de un Gazetteer, que corresponda a las categorías de clasificación, tiene un peso que permite dar un valor en porcentaje de su representación dentro de la clasificación. En el documento se brinda detalle de cómo se definen las categorías y del ciclo de vida bajo el cual se utiliza la herramienta GATE.

El desarrollo de la interfaz gráfica se realiza utilizando tecnología java y en el texto se brinda detalle de cómo se integra con GATE. En resumen, se cuenta con una aplicación denominada JAVA APP que usa los recursos de una aplicación denominada GATE APP. En el diagrama de secuencia se detallan los tipos de datos que cada una de las dos aplicaciones procesa para llegar al resultado de la clasificación.

Como parte del diseño de la herramienta se realiza verificación haciendo comparaciones contra las herramientas IBM Bluemix y Laxalytics. De igual forma se realiza una verificación del

reconocimiento de entidades comparando los resultados del prototipo contra una tarea de etiquetamiento de una muestra de las historias, las mediciones de esta comparación se realizan mediante la funcionalidad de comparación de diferencias que brinda GATE - Developer.

La evaluación mediante TAM se realiza en entrevistas con expertos en el tema. Para lo cual los evaluadores son: un profesor asociado a la Pontificia Universidad Javeriana, un funcionario de la Defensoría del Pueblo y una funcionaria de la Unidad para las víctimas.

Luego del proceso de evaluación del prototipo se encuentra: 1. En cuanto a utilidad tiene validez para las entidades del estado, pues la clasificación orientada a categorías apoya el proceso de análisis imparcial de la información prometiendo generar informes en menor tiempo de lo que actualmente toma. 2. El método con el cual se definen las categorías y el requerimiento de asignar pesos a las palabras en diccionarios es aceptado por los evaluadores. 3. En comparación con otras herramientas en el mercado, tiene ventajas por ser de licencia pública y por permitir definir categorías verdaderamente relevantes para el contexto. 4. Las metodologías utilizadas fueron adecuadas, pues de manera iterativa condujeron al diseño y desarrollo de un artefacto de software que cumple con los objetivos planteados y que es útil para la solución de un problema de la vida real.

De igual forma la evaluación también arrojo oportunidades de mejora que pueden ser una entrada para trabajos futuros. En ellos se tiene: 1. Puede mejorar en elementos de visualización de resultados, de tal manera que se puedan observar tendencias con facilidad. 2. Puede mejorar en mecanismos de control para aquellas historias ausentes de información específica para lograr la clasificación deseada. 3. Puede mejorar en el tratamiento de múltiples historias simultáneamente, de tal manera que los resultados se den individualmente y no como un conjunto. 4. Puede mejorar si se incluyen un mecanismo de consulta que pueda destacar historias de acuerdo a una combinación de categorías deseada.

INTRODUCCIÓN

Los documentos de tipo historias o narrativas se definen como información no estructurada, valiosa en contenido dentro de cualquier área de conocimiento. No contar con limitantes de ingresar datos específicos siguiendo un formulario, permite que esta información esté compuesta de detalles valiosos de tiempo, personas, ubicaciones, eventos entre otros.

En Colombia, entidades como la Defensoría del Pueblo, el Centro Nacional de Memoria Histórica o la Unidad para las víctimas han brindado diferentes informes de la situación actual de las víctimas que el conflicto sociopolítico ha dejado. Dichos informes y la base para su constitución son por completo información tipo texto no estructurada. De acuerdo con las publicaciones el mayor insumo para estos informes son historias o narrativas asociadas a las víctimas. No se han logrado establecer las dimensiones reales de las víctimas que la guerra ha dejado en este país, lo cual confirma que la información continúa en crecimiento.

Esta situación abre un campo de estudio importante para el desarrollo de sistemas de información que impacten positivamente el desempeño en análisis de toda la información no estructurada que se produce a partir de las historias de víctimas del conflicto sociopolítico. En consecuencia, este trabajo plantea utilizar técnicas de preparación de textos que son relevantes dentro del contexto.

El principal aporte de este trabajo será demostrar la validez que tiene un prototipo de clasificación supervisada de historias asociadas a víctimas de la violencia, desarrollado bajo lineamientos de las metodologías Ciencia basada en el diseño y CRISP-DM. Abriendo un espacio importante a futuros trabajos relacionados con el análisis de este tipo de información no estructurada, que es de gran importancia para entidades como la Defensoría del Pueblo o la Unidad para las víctimas.

Este documento se encuentra distribuido en cinco secciones así: 1. descripción general, en donde se describe la problemática que nos motiva a realizar el presente trabajo; 2. descripción del proyecto, en donde se plantea con claridad el alcance del proyecto a partir de los objetivos, metodología utilizada para dar alcance a los objetivos, fases de desarrollo del artefacto y demostración y evaluación del artefacto desarrollado; 3. marco teórico, base teórica para el desarrollo del trabajo; 4. trabajos relacionados, otros trabajos asociados a lo desarrollado para el artefacto, que tienen componentes de aplicabilidad y limitaciones; y 5. modelo de análisis de texto, derivado a partir de la aplicabilidad dada a los trabajos relacionados.

1. DESCRIPCIÓN GENERAL

En esta sección se brinda detalle de los elementos que han motivado la constitución de este trabajo. Se evidencia que al igual que otros proyectos o ideales, se encuentra enmarcado en un deseo de aportar a las labores de construcción de paz.

1.1. Oportunidad y problemática

Cada vez más la construcción de escenarios de paz en el mundo se ve acompañada del soporte de tecnología. En building peace, foro para la paz y la seguridad, se muestra cómo diferentes grupos tienen importantes propuestas, como Stanford Peace Innovation Lab, quienes han identificado tecnologías para la paz, económicas y ubicuas; proponen, por ejemplo, la posibilidad de cuantificar la paz utilizando Facebook para hallar conexiones que se presentan entre personas de diferentes zonas de guerra [2].

Las tecnologías deben ir alineadas a los contenidos relacionados con la construcción de paz. El International Storytelling Center, argumenta que las historias son un elemento importante en las relaciones interpersonales. De acuerdo con su presidenta, las historias o narrativas no contribuyen por sí mismas a la paz — es cómo son empleadas y con qué propósito [3]. Actualmente en Colombia la Defensoría del Pueblo, bajo el proyecto Narrativas Visibles, lleva a cabo "procesos imparciales de construcción, recolección y divulgación, de testimonios de familiares de víctimas y sobrevivientes de la violencia sociopolítica del país". La información que se recolecta es creciente y valiosa en su contenido, teniendo en cuenta que las declaraciones son un requisito de acceso al registro único de víctimas [1]. El análisis de estas narrativas y su utilidad en la construcción de paz se pueden potenciar con el uso de las TIC.

Existe la oportunidad de apropiación de las TIC dentro de la situación actual del país. Como lo proponen Bocanegra et al. [4], las TIC brindan herramientas como blogs, wikis, diferentes soluciones de audio y conferencias, redes sociales y muchas otras, que invitan a la reflexión, el diálogo y la acción. En el presente trabajo, analizar las historias invita a la reflexión, cuando se entiende y se caracterizan las víctimas; invita al diálogo, cuando se hallan elementos comunes entre diferentes protagonistas; e invita a la acción, si al caracterizar la realidad se pueden distribuir y asignar recursos o generar proyectos teniendo en cuenta su composición.

Con la diversidad y cantidad de historias, el registro único para las víctimas, la necesidad de análisis imparciales de la información, entre otros posibles retos; resulta útil hallar un medio apropiado, automático o supervisado, de clasificar información. Esto permitirá estudiar grandes cantidades de texto a costos bajos o disminuir las posiciones particulares asociadas al pensamiento natural de los seres humanos, tal como lo mencionan Quinn et al. [5].

De acuerdo con Grimmer y Stewart [6], quienes han estudiado el análisis de textos políticos, es apropiado emplear técnicas de clasificación de textos supervisadas si se conocen con antelación las categorías dentro de las cuales serán clasificados los textos. En algunos trabajos relacionados vemos cómo se clasifican cuentos infantiles en las categorías de fábula, cuento popular y leyenda [7] o cómo se identifican las categorías de las historias dentro un conjunto cuentos populares [8]. En el caso de historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia, en primer lugar, se deberían identificar dichas categorías y luego encontrar un método apropiado para realizar la clasificación.

El rigor del presente trabajo conlleva a emplear técnicas de preparación de textos, algunas de ellas descritas por Pawar y Gawande [9], Berger et al. [10], Ponte y Croft [11] y Wiedemann [12] y su relevancia aborda la necesidad de diseñar e implementar artefactos de software que apoyen la clasificación de innumerables textos que provienen de las víctimas del conflicto sociopolítico de Colombia, promoviendo así análisis imparciales de las narrativas, brindando alternativas para asignación de recursos, detectando patrones que apoyen reconciliación y otras acciones en el posconflicto, entre otros múltiples beneficios.

2. DESCRIPCIÓN DEL PROYECTO

Esta sección permite evidenciar alcance pactado para este trabajo y su cumplimento. Se brinda detalle del rigor metodológico con el cual se han desarrollado las diferentes etapas y se muestran resultados obtenidos a partir de la comunicación de un prototipo de software desarrollado.

2.1. Objetivos

2.1.1. Objetivo general

Identificar un modelo supervisado de clasificación de textos para historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia que contribuya con la ejecución análisis imparciales.

2.1.2. Objetivos específicos

- Seleccionar a partir del estado del arte técnicas de preparación de texto (segmentación, clasificación, agrupamiento, entre otros) que sean aplicables a este contexto.
- Diseñar la implementación de las técnicas de preparación de texto seleccionadas, así como los artefactos necesarios para esta labor.
- Construir los artefactos necesarios para implementar las técnicas seleccionadas de acuerdo con el diseño de implementación.
- Verificar los artefactos implementados mediante la clasificación de un grupo de historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia.
- Validar mediante evaluación de expertos de la Defensoría del Pueblo la contribución de este trabajo en las actividades de análisis imparciales sobre el tipo de historias analizadas.

2.2. Metodología

Con el objetivo de dar solución a un problema relevante, como lo es contribuir con la ejecución de análisis imparciales de la información proveniente de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia, se abordó como metodología de investigación: Ciencia Basada en el Diseño. Esta es una metodología útil para análisis de problemas aún no resueltos en un ambiente del mundo real y su resolución de una manera novedosa y rigurosa a través del diseño de artefactos [13].

El diseño y la construcción artefactos de software contribuyen a la base del conocimiento y mejoran la efectividad y eficacia de las organizaciones que los utilizan [14], aspectos de la metodología convenientes para abordar el problema expuesto. En su cumplimiento, el desarrollo de este trabajo se guía por un proceso de investigación en sistemas de información, modelo

por fases [15], que alineado con los objetivos benefició la ejecución. Se cumplieron ciclos entre las fases hasta alcanzar los objetivos propuestos, característica propia de la metodología ciencia basada en el diseño.

En concreto, en cada una de las fases se cumplen metas de la siguiente manera: Fase 1, identificación de la motivación; Fase 2, definición de objetivos por cumplir; Fase 3, ejecución iterativa de diseño y desarrollo de artefactos; y Fase 4, cumple con verificación, evaluación y comunicación de resultados.

Dado que el potencial de este trabajo incluye la implementación de técnicas de preparación de textos para las historias en contexto, las fases se apoyan en la metodología estándar más referenciada en minería de datos [16] CRISP-DM.

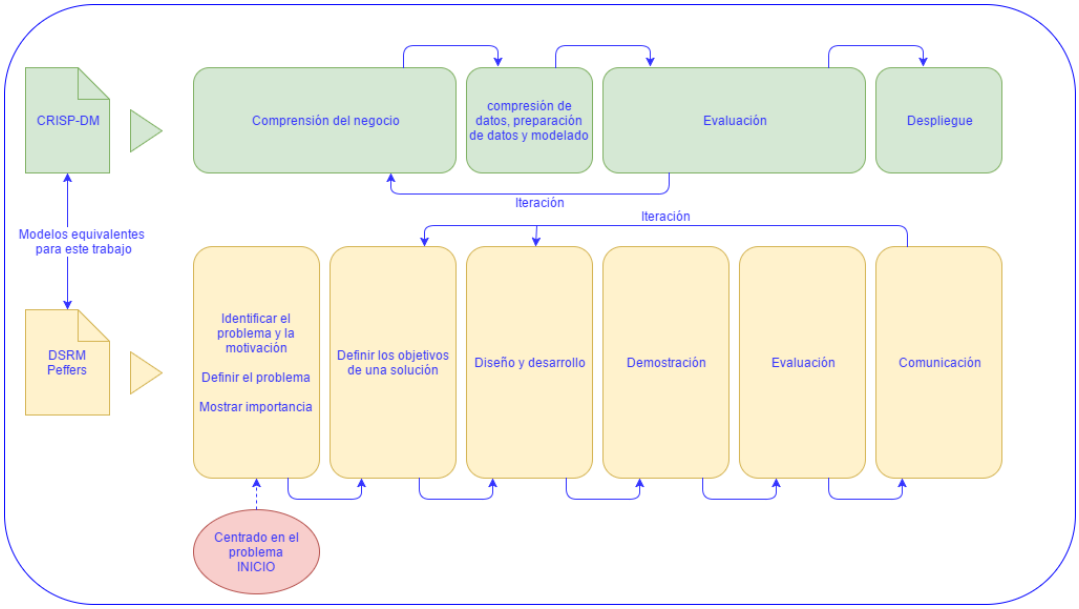


Figura 1 Diagrama metodologías utilizadas (DSRM y CRIP-DM)

En la siguiente tabla se detallan los objetivos y su alineación con las diferentes fases ejecutadas.

Tabla 1Alineación de objetivos a la metodología

Objetivos	Fases según Peffers et al. [15]
1. Seleccionar a partir del estado del arte técnicas de preparación de texto (segmentación, clasificación, agrupamiento, entre otros) que sean aplicables a este contexto.	1. Identificación del problema y motivación 2. Definición de los objetivos (Las fases 1 y 2 son equivalentes a "comprensión del negocio" en CRISP-DM)
2. Diseñar la implementación de las técnicas de preparación de texto seleccionadas, así como los artefactos necesarios para esta labor. 3. Construir los artefactos necesarios para implementar las técnicas seleccionadas de acuerdo con el diseño de implementación.	3. Diseño y desarrollo (La fase 3 es equivalente a "compresión de datos, preparación de datos y modelado" en CRISP-DM)
4. Verificar los artefactos implementados mediante la clasificación de un grupo de historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia. 5. Validar mediante evaluación de expertos de la Defensoría del Pueblo la contribución de este trabajo en las actividades de análisis imparciales sobre el tipo de historias analizadas.	4. Demostración 5. Evaluación (Las fases 4 y 5 son equivalentes a "evaluación" en CRISP-DM) 6. Comunicación (La fase 6 es equivalente a "despliegue" en CRISP-DM)

2.3. Fase de desarrollo

2.3.1. Compresión de datos

El problema relevante identificado, que puede ser solucionado mediante el desarrollo de un artefacto de sistemas de información, es la clasificación de historias provenientes de familiares de víctimas y sobrevivientes del conflicto sociopolítico de Colombia.

Para el desarrollo de esta fase de compresión se realizaron las siguientes actividades:

- Se identifican las historias. Dentro del desarrollo del trabajo y las diferentes iteraciones se encontraron dos fuentes de información pertinentes para el presente trabajo: Centro Nacional de Memoria Histórica, en su documento "¡Basta ya! Colombia: memorias de guerra y dignidad" [17], y Defensoría del Pueblo, en su programa Narrativas Visibles.
- Identificación de trabajos relacionados con técnicas de preparación de textos.
- Se identifica que por las características de los datos es viable realizar un análisis de texto orientado al conocimiento[18].

Los reportes identificados dentro de la metodología como cumplimiento a la fase de comprensión de los datos:

Tabla 2 Reporte de comprensión de los datos

Reporte	Resultado
Colección inicial de datos	En total se consideran disponibles para este trabajo 56 testimonios en documento "Hasta Cuándo", 96 testimonios en documento "La historia detrás de estos ojos", 11 relatos en "La historia no concebida"
Descripción de datos	De los datos descritos como la colección inicial se sabe que la longitud de cada historia es variable, las historias tienen un margen de edición ya que provienen de documentos públicos, luego se encuentran correctamente redactadas.
Exploración de datos	Dentro de las historias se observa que no solo se encuentra un personaje, una localización, un tiempo o delito asociado. Esto comprende un reto para definir las categorías dentro de las cuales se puede realizar la clasificación.
Calidad de los datos	Es una debilidad de este trabajo la calidad de los datos tomados para su ejecución. Está definido que la recolección de datos se encuentra por fuera del alcance y se brinda voto de validez a partir de la labor realizada por instituciones públicas al recolector y publicar los mismos. Con el fin de validar en mayor detalle la calidad de los datos se da respuesta a las preguntas sugeridas por la metodología CRISP-DM en la Tabla 3.

Tabla 3 Validación de calidad de los datos

Pregunta	Respuesta
¿Se encuentran completos los datos?	Como fase inicial de este trabajo se cuenta con historias en los documentos del programa de la Defensoría del Pueblo, Narrativas Visibles. Se toman muestras de estos textos para el desarrollo del artefacto que resuelve el problema planteado, lo cual es suficiente para el alcance de este trabajo. Únicamente se usan datos ya publicados por las instituciones públicas y en entrevistas realizadas a los expertos se observa que hay gran parte de la información que no se encuentra digitalizada o que no es de dominio público.
¿Los datos son correctos?	Desde la perspectiva de este trabajo los datos se asumen correctos, ya que solo se evalúa el tipo de texto y no el contenido de los mismos. La actividad de recolección de datos y su validación ha sido una labor de las instituciones públicas. Las muestras tomadas para este trabajo provienen únicamente de dominio público.
¿Hay valores perdidos en los datos?	Las categorías dentro de las cuales se clasifican las historias son definidas a partir de la interpretación de información de dominio público. La información completa dentro de cada una de las historias no es evaluada, para el desarrollo del artefacto no se identifican valores perdidos, en cuanto al contenido de las historias se refiere. En evaluación con expertos se evidencia que el número total de historias disponibles no es de dominio público, que las categorías de clasificación con las que se realiza este trabajo pueden ser redefinidas y que hay historias que por falta de información pueden no permitir una identificación clara dentro de una categoría respectiva. Está dentro del trabajo futuro extender el prototipo para que se puedan suplir los vacíos a los que conlleva la respuesta a esta pregunta.

2.3.2. Preparación de datos

En busca de una limpieza de los datos, las historias en estudio fueron manualmente preparadas en documentos de texto individuales por historia, remitiéndonos a la información que se encuentra formato pdf y que es publicada como parte del programa Narrativas Visibles.

No se consideran actividades de integración o unión de datos, ya que la fuente información es única y no pretende ser extendida con otros textos relacionados.

Es de recordar que este proyecto se desarrolla empleando la herramienta GATE y que el conjunto de historias individuales se convierte en la base de recursos de lenguaje. Son agrupadas dentro un corpus para utilizar recursos de procesamiento como parte del análisis del texto.

Como parte de las iteraciones realizadas se encontró que es útil preparar los datos realizando una traducción al idioma inglés, ya que muchas técnicas de procesamiento de texto se encuentran bien desarrolladas para dicho idioma. Sin embargo, su utilidad se desvirtúa cuando se realiza análisis orientado al conocimiento, por lo tanto, esta parte de preparación fue descartada.

2.3.3. Modelado

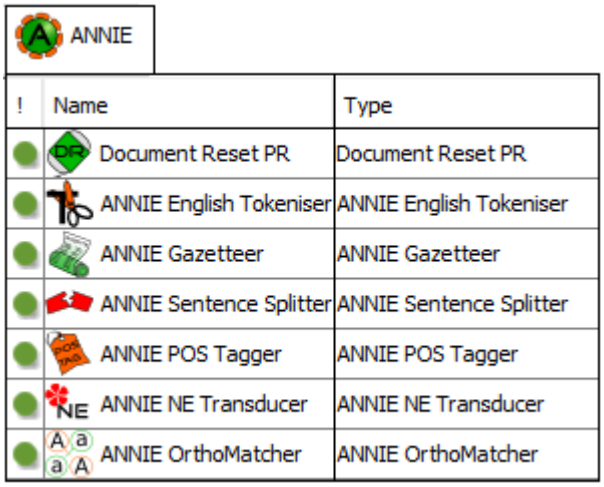
El diseño y desarrollo del artefacto se realizó de forma iterativa hasta alcanzar el objetivo de implementar técnicas de preparación de textos.

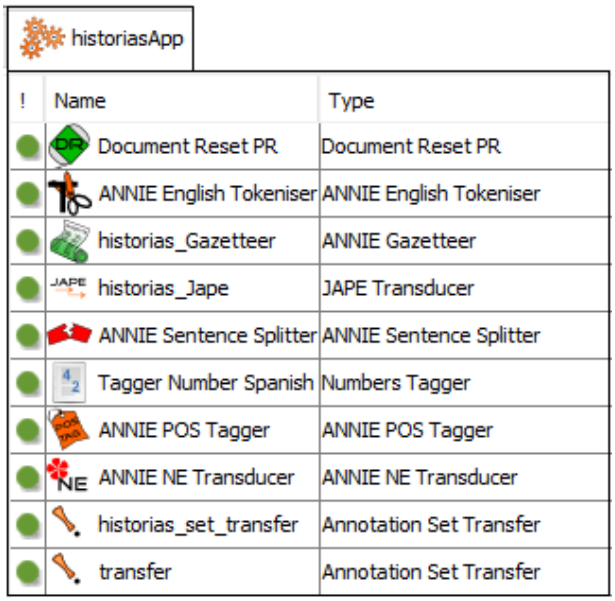
Para el desarrollo de esta fase se realizaron las siguientes actividades:

- Análisis de la estructura de las historias: Se realizó seguimiento al tipo de textos encontrados en los documentos publicados por la Defensoría del Pueblo y se encontró que el contenido relevante no se encuentra estructurado. Así mismo, realizando análisis del documento "¡Basta ya! Colombia: memorias de guerra y dignidad"[17], se encontró una categorización de modalidades de la violencia perfectamente aplicable a las historias publicadas por la Defensoría del Pueblo.
- Selección de los métodos a utilizar para preparación de textos.
- Selección de la herramienta para implementación de técnicas de preparación de textos.
- Desarrollo y verificación, de forma iterativa, de los artefactos necesarios para la implementación de técnicas de preparación de textos aplicadas a las historias del contexto.

En cumplimiento a la fase de modelado, dentro de la metodología se obtienen los siguientes resultados a las tareas de la fase:

Tabla 4 Reporte modelado

Tarea	Resultado
Seleccionar técnica de modelado	Análisis de texto orientado al conocimiento. Se apoya en la definición de diccionarios por categorías de interés, reconocimiento de entidades a partir de lo definido en diccionarios y reglas selección para identificación de los segmentos de texto con información de interés.
Generar diseño de prueba	Como diseño de prueba se utiliza directamente la aplicación ANNIE que es definida por defecto dentro de la herramienta GATE. Se determina la viabilidad de reconocer entidades como ciudades asociadas a las historias.  **Figura 2 Aplicación ANNIE**
Construir el modelo	1. Crear aplicación de tipo Corpus Pipeline en GATE Developer 2. Definir recursos de procesamiento que tengan en cuenta los artefactos propios de las historias de víctimas ❖ 2.1. Gazetteer o diccionarios: Listas de términos de referencia para realizar extracción de la información. Las listas se encuentran definidas por categorías para referencias propias de este contexto y se complementan con listas comunes a proyectos de este tipo (nombres, ciudades, números, meses, entre otras) ❖ 2.2. JAPE: Lenguaje para representación de la información extraída. Las reglas creadas se ejecutan de igual manera en forma de pipeline.

	❖ 2.3. POS Tagger: El etiquetamiento fue descartado parcialmente por tratarse de idioma español y no encontrar mayor beneficio al realizar una traducción para poder utilizar este recurso. ❖ 2.4. Number Tagger: Plugin de propósito general utilizado para la identificación de los números. ❖ 2.5. Annotation set transfer: Utilizado para extraer entidades ya conocidas y transferirlas a una agrupación personalizada para el proyecto.  **Figura 3 Pipeline prototipo**
Evaluar el modelo	El modelo y prototipo se evalúan desde dos perspectivas relevancia y validez. En el mercado se encuentran herramientas que permiten realizar análisis a los textos, apoyando las labores de análisis cualitativo; sobre dos de ellas se ha realizado una comparación con el fin de determinar si este trabajo tiene impactos positivos que confirman la importancia de su implementación y su relevancia. De igual forma se ha realizado una comparación de los resultados de entidades reconocidas a partir de una labor manual de etiquetamiento de las muestras de historias de víctimas de la violencia. Se utiliza la funcionalidad de realizar revisión de diferencias, que se encuentra en el ambiente

GATE Developer, para generar una cuantificación de los resultados.

Las herramientas utilizadas como punto de comparación de la funcionalidad son: Plataforma IBM Bluemix, empleando el servicio *IBM Watson*, y Laxalytics, empleando *Text Analytics Demo* y *Semantria for Excel*.

Pruebas con IBM Bluemix:

Para las pruebas se crea una cuenta temporal (30 días), ya que no se trata de software libre y se deben saldar los costos pertinentes para dar utilidad sin restricción.

Se elaboró una aplicación AlchemyAPI, que permite gestionar conexión a funcionalidades de procesamiento natural del lenguaje para el análisis de texto, como las que se proveen mediante *IBM Watson.*

Dentro de las pruebas realizadas se encuentra que el servicio Node-RED brinda un espacio de trabajo amigables al usuario. La interfaz gráfica es intuitiva y permite recrear flujos de actividades con facilidad.

El siguiente es el flujo recreado para realizar comparación de resultados:

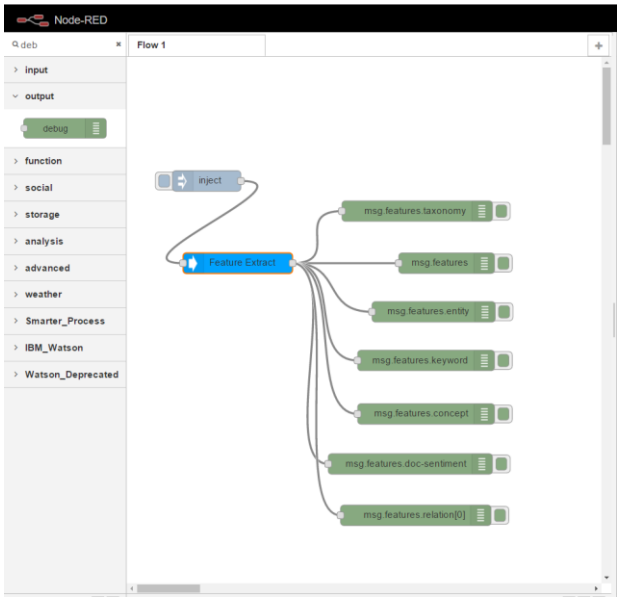


Figura 4 Configuración Node-RED

Este flujo fue probado con la historia de una víctima de la violencia y

como resultado se obtuvo una respuesta en formato JSon para cada una de las características encontradas en el texto. La herramienta arroja resultados seleccionados a partir de la relevancia dentro del texto, es de aclarar que no es claro a partir de qué se realiza el ponderado de la relevancia.

De igual forma, es importante resaltar que no fue posible identificar la manera de cómo se podían conocer los resultados completos de las características extraídas.

La imagen muestra la respuesta al componente de salida msg.features.keyword

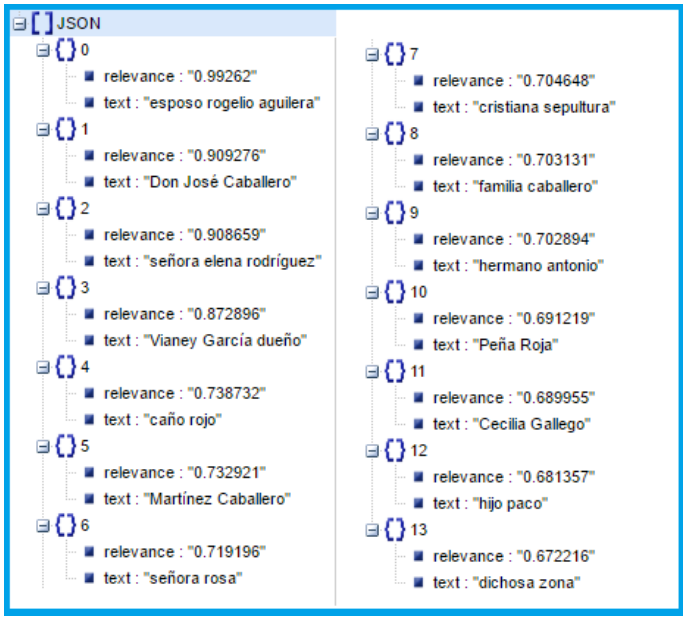


Figura 5 JSon resultados msg.featuras.keyword

Los resultados se observan en ventana de debug como se ilustra en la siguiente imagen:

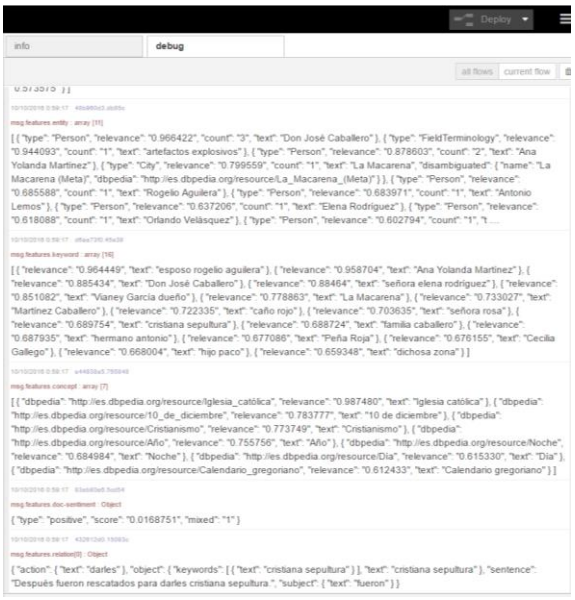


Figura 6 Resultados Node-RED

Dentro de las pruebas realizadas no se identificó una forma de alimentar las categorías de interés, con el fin de realizar análisis orientado a categorías definidas por un usuario.

Pruebas con Lexalytics:

La prueba inicial se realiza con *Text Analytics Demo*, el cual cuenta con una interfaz muy básica e intuitiva. Las posibilidades de error en la utilización del demo son mínimas, ya que la funcionalidad es reducida y cuenta con elementos como mensajes para orientar al usuario.

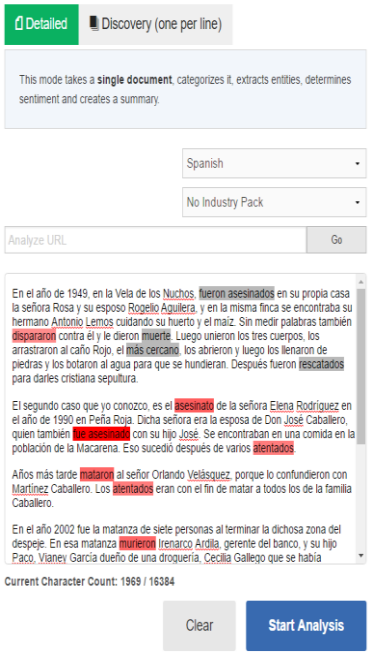


Figura 7 Text Analytics Demo

Se observa la identificación de términos asociados al tema “Crimen” que realiza la herramienta con alto grado de asertividad. La herramienta encuentra una categoría asociada a la palabra por su recurrencia y la de sus sinónimos dentro del texto. Es lo que se denomina una asociación tipo Cluster.

Las siguientes dos imágenes corresponden a los resultados obtenidos a partir de esta herramienta. En donde se evidencia que las entidades son las personas reconocidas dentro del texto y las categorías son las palabras generadas a partir de aplicar asociación por Cluster.

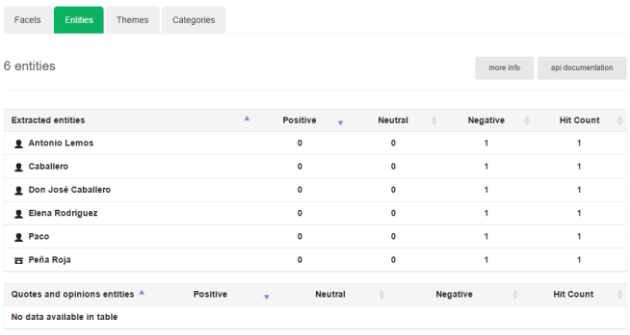


Figura 8 Entidades - Text Analytics Demo

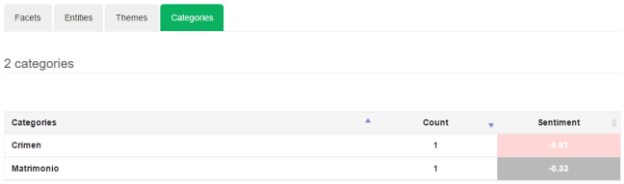
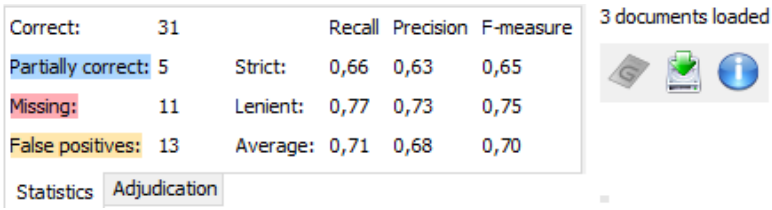


Figura 9 Categorías - Text Analytics Demo

No se evidencia fácilmente una forma de subdividir la categoría, de tal manera que se pueda brindar un mayor detalle a la clasificación de la historia.

Resultados del desempeño del prototipo

Comparación en GATE contra un etiquetado manual: Consiste en comparar la misma historia etiquetada por el pipeline del prototipo contra la etiquetada manualmente.



En promedio las unidades de Recall y Precision superan el 60%, por lo cual se procede a aceptar el prototipo para su validación con expertos.

2.3.4. Prototipo

El desarrollo del prototipo se realizó utilizando tecnología Java. Se trata de una aplicación Java standalone, administrada como un proyecto Maven.

Herramientas de desarrollo JAVA APP:

- NetBeans IDE 8.0.2
- JDK 1.7

Herramientas de desarrollo GATE APP:

- GATE Developer 8.1
- Gazetteer personalizados para este trabajo
- Archivos JAPE personalizados para este trabajo.
- ANNIE Plugin
- OpenNLP Plugin
- Tagger Numbers Plugin

Presentación prototipo:

Prototipo de única interfaz, representada en la Figura 10, que permite cargar los archivos por analizar y visualizar los resultados de reconocimiento de entidades en dos tablas, la primera tabla ilustra un resumen con porcentajes de cada entidad respecto al total de las identificadas en el texto y la segunda tabla muestra el detalle de cada una de las entidades reconocidas y, como parte de la funcionalidad supervisada, permite modificar los pesos que han resultado para cada entidad.

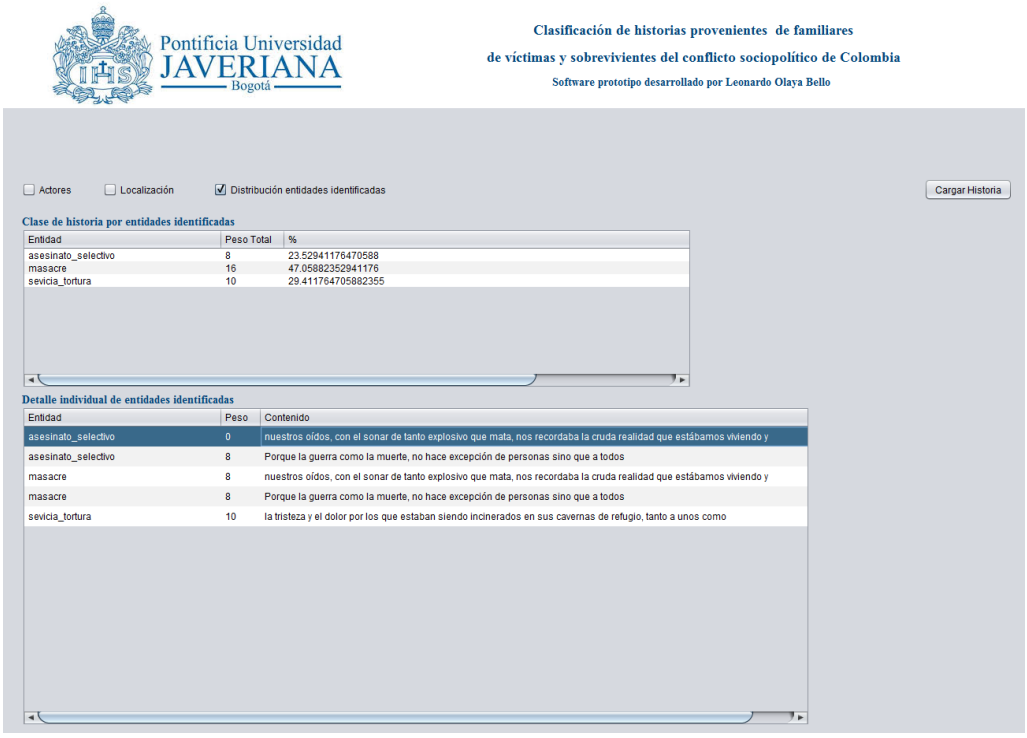


Figura 10 Presentación del prototipo

Diagrama de secuencia:

El siguiente diagrama, Figura 11, ilustra el flujo de procesamiento de una historia, desde su solicitud de carga al sistema hasta la visualización del resultado.

Detalla cómo JAVA APP es el medio de interacción entre el usuario y GATE APP. En la proximidad con el usuario cumple con la responsabilidad de recibir archivos de texto, conformar un GATE Corpus y entregar resultados en presentación de tablas a partir de la lectura archivos xml. En la proximidad con GATE APP cumple con la responsabilidad de entregar un GATE Corpus y recibir resultados de entidades reconocidas, resultados que son leídos y almacenados como archivos XML.

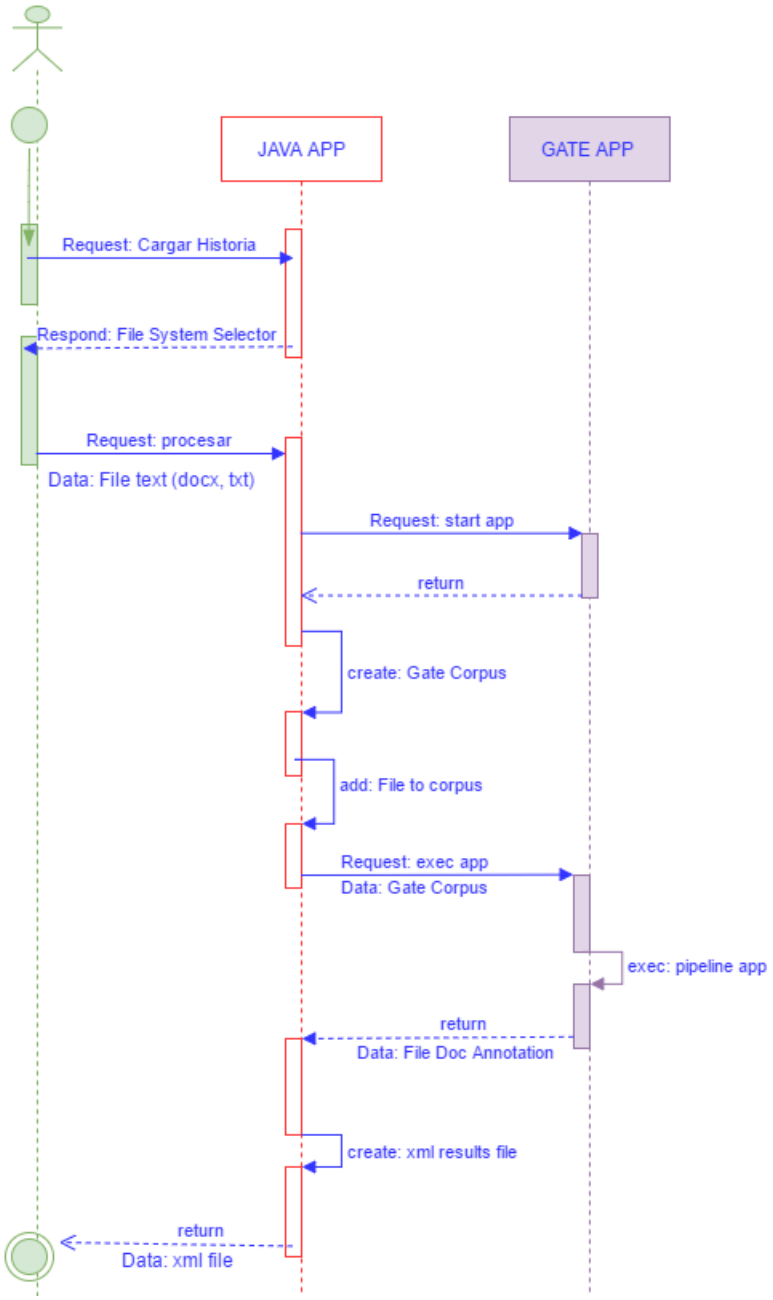


Figura 11 Diagrama de secuencia

Diagrama de procesos:

El desarrollo se específica dentro de dos procesos independientes, como se ilustra en Figura 12. El primero con responsabilidad de administrar archivos por analizar y procesar los resultados obtenidos, para su visualización, en la Figura 12 se identifica como un proceso de control de JAVA APP. El segundo con responsabilidad de ejecutar el pipeline definido para analizar los textos, en la Figura 12 se identifica como GATE APP.

El diagrama ilustra la utilización que hace JAVA APP de GATE APP. Esto se da porque se está utilizando el entorno GATE Developer para realizar la codificación del pipeline de análisis, el resultado de esta codificación es una aplicación ejecutable independiente, que es invocada por las clases definidas en JAVA APP.

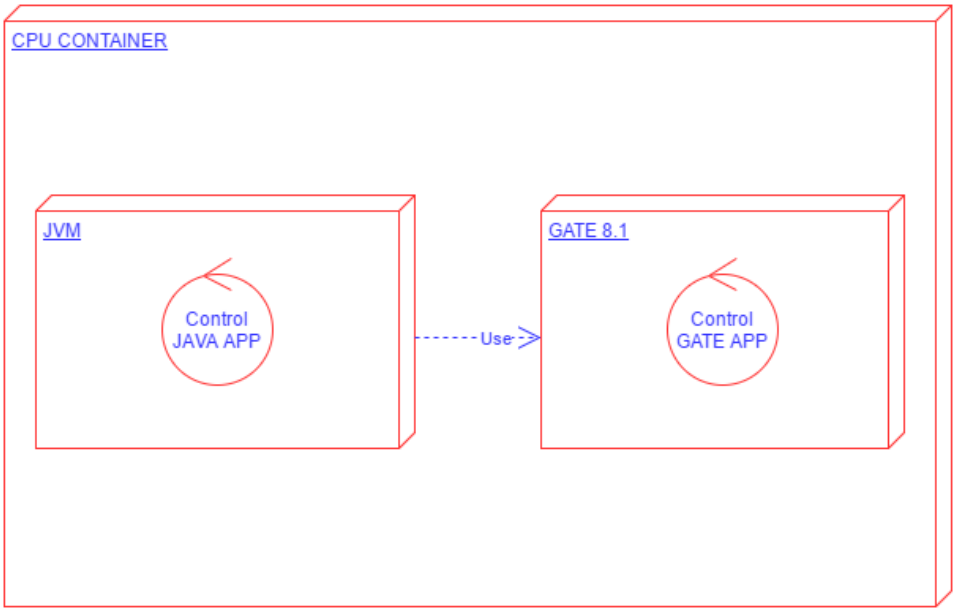


Figura 12 Diagrama de procesos

2.4. Demostración y evaluación

2.4.1. Modelo de evaluación

El modelo empleado para esta evaluación es Technology Acceptance Model (TAM), la percepción de utilidad y de fácil utilización son el objetivo de este modelo [19].

De forma consecuente al modelo, las actividades realizadas han tenido un enfoque en el usuario y no en el artefacto, facilitando el objetivo de evaluar la utilidad, ya que se pretende solucionar un problema de la vida real.

Se realizaron entrevistas a tres expertos, considerados como potenciales usuarios del artefacto desarrollado en este trabajo. Las entrevistas pretendían conducir el trabajo por el campo en el que se encuentre la mayor relevancia y determinar así mismo el grado de aceptación y las oportunidades de mejora.

Las observaciones de aceptación por parte de los expertos son consideradas como validación de utilidad para resolver el problema y las observaciones de mejora son consideradas como oportunidades de una próxima iteración en ciclo de diseño.

2.4.1. Resultados de evaluación con expertos:

Tabla 5 Resultado evaluador 1

Evaluador 1	
Nombre	Wilson López López
Cargo	Profesor asociado a la Pontificia Universidad Javeriana Facultad de Psicología
Aceptación	En las entrevistas con evaluador se observó un alto grado de aceptación en el tema planteado para este trabajo, destacado por aportes importantes en cuanto a campos de relevancia y fuentes de información representativas. Su aporte nos conduce a la Unidad de atención para las víctimas, como entidad en la cual este trabajo es relevante, y a la fuente de información del Centro Nacional de Memoria Histórica, como entidad con un informe detallado que nos permite identificar categorías sobre las cuales enfocar el análisis de las historias.

Oportunidad de mejora	❖ Ampliar el concepto de clasificación hasta apoyar procesos de observación de víctimas, de tal manera que se puedan sugerir tratamientos psicológicos de acuerdo con el grado de victimización. ❖ Identificar el rastro que han tenido las víctimas en el territorio nacional, de tal manera que se pueda dar un mejor uso de los recursos del estado o de cualquier ONG para apoyo en el tratamiento a su condición.

Tabla 6 Resultado evaluador 2

Evaluador 2	
Nombre:	Wilson Herney Chavarro Jiménez
Cargo:	Asesor nacional de la defensoría delegada para la orientación y asesoría a las víctimas del conflicto armado interno. Responsable del proyecto de memoria histórica Narrativas Visibles
Aceptación	En la entrevista con evaluador se observó un nivel de aceptación intermedio, destacada por un interés en la presentación de reportes a partir de la información que ya tienen dentro de su documentación. Se resalta la necesidad de contar con desarrollos de este tipo, ya que la cantidad de información que manejan es alta y no cuentan con sistemas de información que apoyen las labores de análisis de la misma información. En la demostración se realizó análisis con una historia del programa Narrativas Visibles, que se encuentra como documento de texto (*.docx). Con esta prueba se evidenció el reconocimiento de las entidades predefinidas en ese momento. El evaluador mostró su interés en cumplir esa misma labor para múltiples historias simultáneamente. Durante la evaluación se explica el método con el cual se definen las categorías y este es aceptable para el evaluador.

Oportunidad de mejora	❖ Ingresar un documento completo en formato pdf, con múltiples historias, y extraer de manera detallada, por cada historia, las entidades reconocidas. ❖ A partir de las entidades reconocidas tener la posibilidad de realizar consultas por combinación de las mismas. Ejemplo: encontrar las historias asociadas a grupos guerrilleros en las que se observen delitos como sevicia y tortura.

Tabla 7 Resultado evaluador 3

Evaluador 3	
Nombre:	Mireya Camacho
Cargo:	Asesora Director General Unidad para las víctimas
Aceptación	En la entrevista con el evaluador se observó un nivel de aceptación alto, en donde se destacó el interés por el análisis imparcial de las historias que ya se tienen de las víctimas. Se relata que la dirección de registro y observaciones para las víctimas cumple una labor importante al realizar el análisis de las historias, en muchos de los casos se deben presentar informes detallados que se pueden apoyar en un sistema de este tipo. Tener la posibilidad de definir categorías de interés y crear diccionarios con listas de palabras asociadas a los mismos es una labor que se puede realizar con expertos de esta entidad gubernamental. Durante la evaluación se explica el método con el cual se definen las categorías y el requerimiento de asignar pesos a palabras dentro de los diccionarios, lo cual resulta aceptable para el evaluador.
Oportunidad de mejora	❖ Visualización de la información. Es importante contar con gráficas de tendencias de las diferentes categorías encontradas dentro de las historias. ❖ Definir claramente variables de control para aquellos casos en los que los resultados se den dentro de la ausencia de información. En

	una historia se pueden presentar casos en los que no están claramente definidos los actores y es necesario mitigar el riesgo de error ante una situación de este tipo.

3. MARCO TEÓRICO

Esta sección brinda una visión general de la base teórica consultada para desarrollar el presente trabajo.

3.1. Análisis de texto asistido por computador

Áreas como humanidades, ciencias sociales, medicina, entre otras, apoyan sus estudios en herramientas y técnicas que asisten el estudio de textos o información no estructurada [12], [20], [21]. Estas herramientas o técnicas permiten realizar análisis cuantitativos y cualitativos de los textos, un ejemplo de esto se puede observar en el trabajo realizado por Borja Orozco et al. [22].

De acuerdo con el estudio realizado por Gregor Wiedemann [12], es representativa para este trabajo la siguiente clasificación de tipos de análisis de textos: 1. Enfoque No Supervisado vs. Supervisado, y dentro de las supervisadas se tienen 2. Estadístico vs. Lingüístico y 3. Deductivo vs. Inductivo.

De igual manera son representativos dos de los enfoques descritos para análisis de texto asistido por computador:

- "Enfoque de análisis de contenido computacional basado en hipótesis, (CCA) proporciona anotaciones de textos automáticamente a través de la observación de ocurrencia de términos. El cual permite la disminución de grandes cantidades de texto mediante la identificación de categorías, definidas por expertos, dentro de los mismos".
- "Enfoque de minería de texto que se esfuerza en la extracción de "significado" mediante la aplicación de modelos estadísticos. Los métodos estadísticos son combinados con conocimiento manualmente codificado con expertos, de tal manera que se puedan identificar patrones dentro de los contenidos".

3.2. Procesamiento natural del lenguaje

Procesamiento natural del lenguaje (NLP) se refiere a un número de tareas computacionales que pueden ser aplicadas para el análisis de textos, por ejemplo la identificación automática de segmentos de texto que proveen conceptos de interés teórico necesarios para realizar análisis cualitativo [18].

Los niveles de análisis lingüísticos que pueden ser realizados mediante técnicas de NLP fueron definidos, en términos del idioma inglés, por Crowston et al.[18] con la siguiente tabla:

Tabla 8 Niveles de análisis lingüísticos

Nivel	Definición y ejemplo
Fonológico	Auditar características del lenguaje: sonidos, tono e inflexiones
Morfológico	El menor nivel de significado lingüísticos, los morfemas. Por ejemplo, prefijos y sufijos que cambian el significado de las palabras.
Léxico	Nivel de palabra. Parte del discurso es una característica del análisis léxico que afecta el significado. Consideremos, por ejemplo, la diferencia de significado entre "*book*" como sustantivo ("*read a book*") y "*book*" como un verbo ("*to book a flight*")
Sintáctico	Significado que deriva de la secuencia de palabras en una frase o frase. Por ejemplo, considere los diferentes significados de "*the man hit the ball*" y "*the ball hit the man*".
Semántico	Definición de los significados de las palabras dentro del contexto, ya sea que el nombre "*bank*", por ejemplo, se refiera a una orilla del río ("*river bank*") o a una institución financiera. El análisis semántico puede tratar con grados de significado dependiendo del contexto.
Discurso	Significado basado en una unidad más grande que una oración, donde el significado de una oración particular es afectado por el texto que lo precede o su colocación dentro de un documento. El análisis del discurso ha llevado a la identificación de géneros de documentos, donde la información se puede encontrar predecible a través de la estructura del documento (introducción, por línea, resultados de la investigación, etc.)
Pragmático	Incorporación del conocimiento mundial para determinar el significado, es decir, connotaciones basadas en la experiencia y entendimientos compartidos. Por ejemplo, entendemos mucho más acerca de los "Third World Countries" que las palabras que los componen solo nos pueden decir.

3.3. Minería de texto

Minería de texto se considera como la actividad de extraer información, a partir de un conjunto extenso de textos, utilizando herramientas de software. Emplea un amplio rango de estadística, aprendizaje de máquina, y técnicas lingüísticas que son asociadas al NLP [23].

Típicamente una actividad de minería de texto cumple con la ejecución en pipeline de una serie de tareas de NLP, que son usadas para dar formato al texto, en preparación para el análisis estadístico o la fase de descubrimiento de patrones. En dichas actividades se destaca el reconocimiento de entidades que es altamente utilizada en áreas de investigación como la medicina. A continuación, una tabla desarrollada por Harpaz et. al.[23], donde se describen las actividades típicas de NLP que son aplicadas en la mayoría de proyectos de minería de texto:

Tabla 9 Actividades típicas de NLP que son aplicadas en minería de texto

Tarea	Descripción
Segmentación	División de un documento a lo largo de los límites de las oraciones y las secciones.
Tokenización	División de las oraciones entre sus partes - palabras y puntos.
Part of speech (POS) tagging	Asignar partes de discurso gramaticales a elementos individuales, ejemplo *"Drug"* es un nombre, *"administer"* es un verbo, *"quickly"* es un adjetivo, *"the"* o "a" son determinantes.
Parsing	Determinar la estructura gramatical de las oraciones y la relación entre los grupos de palabras que forman conjuntamente frases nominales, frases verbales, cláusulas, etc. El análisis superficial, a menudo utilizado en lugar del análisis profundo, sólo identifica los constituyentes (por ejemplo, sinónimos) pero no la estructura interna de la oración.
Reconocimiento de entidades (NER)	Identificar términos o frases de interés ("entidades") en el texto. NER puede ir más allá de simplemente reconocer los términos para también categorizar, normalizar y asignarlos a vocabularios estandarizados.
Detección de negaciones	Determinar si una entidad con nombre está presente o ausente, Ejemplo" Paciente no presenta síntomas de …", "paciente fue descartado por infarto de miocardio"
Desambiguación del sentido de palabra (WSD)	Determinar qué sentido de un homógrafo (palabras con idénticas ortografías, pero con significados diferentes) es apropiado en el contexto de la oración

Inferencia temporal	Establecimiento de un orden temporal de eventos a partir del texto, Ejemplo “Acontecimiento adverso ocurrió después de la prescripción de la droga"
Detección de relación	Determinar si dos o más entidades con nombre reconocidas en el texto forman relaciones específicas, Ejemplo “El fármaco A trata la enfermedad B”, “el fármaco A induce la enfermedad B”

3.4. Reconocimiento de entidades

Reconocimiento de entidades es una tarea asociada al procesamiento del lenguaje, que busca extraer información dentro de un texto a partir de una serie de datos ya conocidos. Dentro de los métodos aplicados para esta labor se encuentra la utilización de Gold estándar o el empleo de los diccionarios [24].

Para Wiedemann [12] el fin de extraer significado de los textos, sobrepasando la actividad de conteo de cadenas de caracteres, soporta la necesidad de realizar inclusión sistemática de análisis de contextos de los cuerpos en estudio. Es una tarea que puede ser realizada mediante el reconocimiento de entidades utilizando diccionarios, los cuales se definen para cada contexto específico y son ampliamente aceptados, teniendo en cuenta que por su naturaleza son estructuras que deben ser mantenidas y sometidas a cambios periódicos. Listas de términos, describiendo categorías de interés, cumplen con el propósito de diccionarios.

3.5. GATE – General architecture for text engineering

GATE es uno de los sistemas más ampliamente utilizados dentro de aquellos de su tipo, con tasas de descarga de diez miles y gran cantidad de usuarios activos en medios académicos y contextos industriales [25]. Con más de 15 años, el software se encuentra activo para su utilización en cualquier tarea computacional asociada al lenguaje humano.

Algunos los aspectos importantes:

- JAPE: Java Annotation Patterns Engine, es una versión de “CPSL – Common Pattern Specification Language”. La gramática de JAPE consiste en un conjunto de fases, de las cuales cada una consiste en un conjunto de patrones/reglas. Las reglas denominadas de mano izquierda (LHS) constituyen la descripción de un patrón dentro de un grupo de anotaciones o etiquetas. Las reglas denominadas de mano derecha (RHS) consisten

en declaraciones de manipulación de una anotación, definen la acción a realizar cuando un patrón se ha detectado dentro del texto gracias a lo definido por una LHS [26].

- ANNIE: Es un ejemplo de un sistema de extracción de información suministrado por GATE y que se encuentra con una aplicación pipeline ya definida, que permite facilitar el trabajo realizado con este sistema para cualquier corpus de interés.
- CREOLE Plugins: Conjunto de recursos de procesamiento que pueden ser administrados dentro de GATE para realizar tareas específicas. De acuerdo con su funcionalidad y diseño, requieren de parametrizaciones iniciales específicas para su correcta ejecución.
- GATE Developer: Interfaz gráfica de GATE que permite diseñar y ejecutar aplicaciones sobre un cuerpo de textos. La principal actividad a diseñar en una aplicación GATE consiste en generar un conjunto de anotaciones sobre textos. Cargar y ver documentos, crear y ver conjuntos de documentos, trabajar con anotaciones, utilizar CREOLE Plugins, son algunas de las tareas que pretende facilitar la utilización de GATE Developer.
- GATE Embedded: Es un framework orientado a objetos (librería de clases) desarrollado en java y disponible bajo GNU Lesser General Public Licence 3.0. Al igual que GATE Developer, permite realizar la debida administración de recursos de lenguaje, procesamiento y visualización.

4. TRABAJOS RELACIONADOS

Como parte del desarrollo de una herramienta de software que permita realizar una clasificación de textos, dando tratamiento al contenido para cumplir con las expectativas propuestas. Por esta razón los trabajos de Cunningham et al. [25] y Dehghan et al. [27] son relevantes, en adelante TR-I y TR-II respectivamente.

Cunningham et al. [25] propone, desde la perspectiva de textos médicos, el ciclo de vida del análisis de este tipo de narrativas utilizando la herramienta GATE y Dehghan et al. [27] muestra dentro de su trabajo qué procesos realizar para el reconocimiento de entidades mediante métodos conducidos por datos y conocimiento, sin ser este el fin principal de su trabajo.

4.1. Ciclo de vida del análisis de texto utilizando GATE

Aplicado en el contexto de la medicina, el TR-I muestra la importancia de la herramienta GATE, "definida como una herramienta ampliamente utilizada para el procesamiento de texto en sistemas de uso frecuente con tasas de descarga anual de decenas de miles y muchos usuarios activos en tanto académica y contextos industriales" [25].

4.1.1. Aplicabilidad

TR-I se encuentra ampliamente relacionado con lo que se realiza a nivel de historias de víctimas de la violencia, pues en ambos casos se trata de información no estructurada que se encuentra como texto. De la mano con lo descrito por TR-I, en un futuro cercano no se ven oportunidades para realizar actividades en las cuales se pueda estructurar la información contenida en las historias de víctimas de la violencia, ya que esto conlleva a implicaciones de costos e inflexibilidad para el manejo de la información.

El aprovechamiento de conocimiento basado en diccionarios es propio de los estudios en documentos asociados a la medicina, pues se requiere utilizar recursos específicos para el dominio.

El planteamiento del ciclo de vida para análisis de texto en TR-I define metódicamente cómo abordar un estudio de este tipo empleando la herramienta GATE. Los pasos descritos son:

- Conformar una colección de textos sobre los cuales se requiere realizar observaciones, lo cual se denomina *"corpus"* o *"collection of corpora"*.
- Desarrollar una descripción estructurada de los aspectos de interés dentro del texto.
- Especificar las tareas de extracción y verificar la especificación (como sugerencia para proyectos pequeños se plantea emplear *GATE Developer*) para marcar manualmente un GOLD ESTÁNDAR.
- Elaborar un prototipo de pipeline para análisis de texto.
- Desplegar y verificar el sistema de análisis.
- Poblar un servidor de índices.

- ❖ Exponer los resultados a usuarios finales.

4.1.2. Limitaciones

TR-I propone el aprovechamiento de recursos de procesamiento diseñados para trabajos asociados a medicina, entre ellos: ABNER (A Biomedical Named Entity Recognizer), MetaMap (maps biomedical text), Gspell biomedical spelling suggestion and correction o BADREX (identifying Biomedical Abbreviations using Dynamic Regular Expressions), por mencionar algunos.

Es claro que el trabajo asociado a los documentos médicos tiene apoyo amplio de la comunidad, como lo sugieren los plugins mencionados. Sin embargo, en el caso de historias de víctimas de la violencia se tienen dos componentes que limitan las alternativas de trabajo y son: primero el idioma y segundo son narrativas en un dominio específico enmarcado en un estricto contexto asociado a la realidad de Colombia.

Se plantea dentro del método la elaboración de un GOLD ESTÁNDAR que no puede ser definido para el trabajo con las historias de las víctimas ya que no hay trabajos de base dentro de este contexto. De igual forma se plantea dentro del trabajo a realizar, el aprovechamiento de reconocimiento de entidades, de tal manera que se permita suplir la desventaja.

4.2. Reconocimiento de entidades mediante métodos conducidos por datos y conocimiento.

El proceso para realizar el tratamiento del texto, de tal manera que sea posible realizar un reconocimiento de entidades, se guía por una serie de acciones ejecutadas en forma de pipeline. Como ejemplo de pipeline en TR-II se realiza el procesamiento de las narrativas, definido mediante una serie de tareas a ejecutar sobre el texto de manera independiente.

4.2.1. Aplicabilidad

TR-II especifica su trabajo mediante la herramienta GATE y define los pasos a ejecutar según su momento así: 1. Pre-procesamiento, 2. Diccionarios y etiquetadores basados en reglas, 3. Etiquetadores base ML, 4. Reconocimiento en segunda fase y 5. Integración.

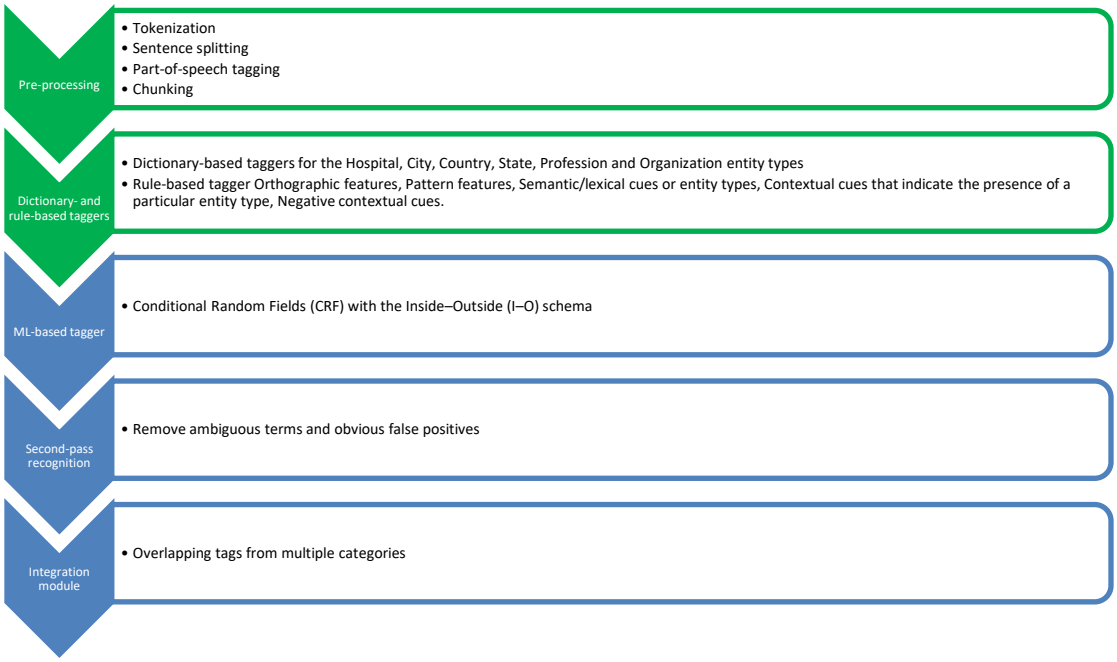


Figura 13 Pasos a ejecutar en procesamiento

De manera complementaria justifica la utilización de diccionarios bajo el nombre análisis dirigido por conocimiento. Para este caso puntual se realiza énfasis dentro de sus resultados por mejorar notablemente los resultados en reconocimiento de entidades como organizaciones (+25%) y profesiones (+56%).

4.2.1. Limitaciones

TR-II se encuentra orientado a la anonimización de las narrativas clínicas, luego el número de categorías definidas para realizar el análisis asistido por conocimiento es mínimo. En el caso específico del trabajo con las narrativas de historias de la violencia y su clasificación, el número de categorías juega un rol importante, pues se incrementa el número de diccionarios y asociación asistida en reglas proporcionalmente al número de categorías necesarias. Por lo tanto, es comparable por completo en metodología para ejecución del análisis, pero no en resultados de identificación de entidades.

5. MODELO DE ANÁLISIS DE TEXTO

5.1. Definición general del modelo

La definición del modelo de análisis de texto se fundamenta en cuatro ítems:

- Ciclo de vida para análisis de texto utilizando la herramienta GATE.
- Diccionarios especializados en el contexto.
- Pre procesamiento de texto.
- Etiquetadores basados en reglas y diccionarios.

En relación con TR-I se abordaron las siguientes acciones:

- En cumplimiento al ciclo de vida en GATE, se encontró que no se encuentra perfectamente definido un Gold estándar asociado a este proyecto. Sin embargo, se plantea la construcción de diccionarios por categorías a estudiar, para lo cual se cumple con las siguientes tareas:
  - Se utiliza la clasificación impartida en el documento "¡Basta Ya!" [17], del Centro Nacional de Memoria Histórica.
  - Se realiza el análisis de las categorías mencionadas dentro del documento y se definen palabras claves asociadas a dichas categorías.

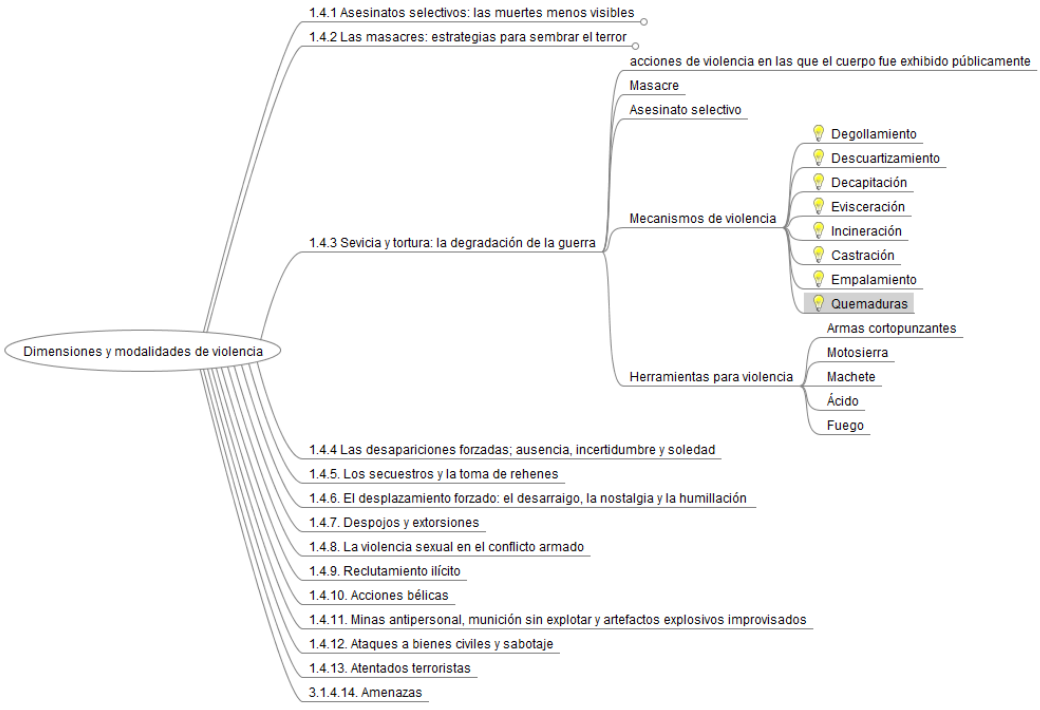


Figura 14 Categorías y palabras claves asociadas

- ➢ Se elabora un diccionario por la categoría a estudiar, cumpliendo así con diccionarios especializados en el contexto.
- ➢ Se brindan pesos a cada una de las palabras asociadas a los diccionarios. Es de resaltar que los pesos pueden ser ajustados en el momento de la supervisión de entidades reconocidas.
- ➢ Se asocia el diccionario al pipeline definido para el procesamiento del texto.

- ❖ Como parte del pre procesamiento se realizan actividades limpieza de entidades sobre el documento, tokenización en palabra y fraccionamiento en oraciones.

Como consecuencia de la ejecución de tareas sobre las narrativas clínicas en TR-II, se extraen las tareas homólogas aplicables sobre narrativas de víctimas de la violencia. Por esta razón, se aplican las dos primeras fases (pre-procesamiento y reglas de etiquetamiento basadas en diccionario) y se ejecutan mediante el apoyo de la herramienta GATE como se especifica en la siguiente imagen, como recursos de procesamiento.

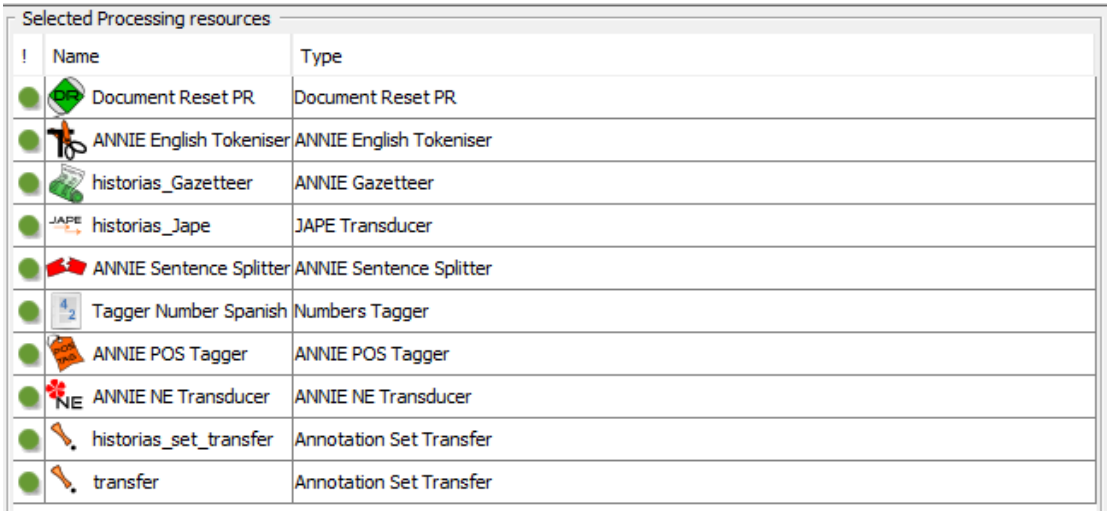


Figura 15 Pipeline prototipo

Donde los recursos de procesamiento se ejecutan en orden de pipeline desde Document Reset PR hasta Annotation Set Transfer.

De manera detallada cado uno de los recursos de procesamiento ejecuta:

- ❖ Document Reset PR: Limpieza del texto para evitar que otros procesos hayan afectado el estado inicial, en cuanto a etiquetas se refiere.
- ❖ ANNIE English Tokeniser: Tokenización del texto en palabras.

- ANNIE Gazetteer: Aplicación de reglas de concordancia mediante la comparación con diccionarios.
- JAPE Transducer: Reglas de asociación para extracción de información a partir de los reconocimientos realizados mediante el paso por diccionarios.
- ANNIE Sentence splitter: Segmentación del texto en oraciones.
- Tagger Number Spanish. Etiquetamiento de números.
- ANNIE POS Tagger: Ejecución de *part of speech tagging.*
- ANNIE NE Transducer: Traducción de reconocimiento de POS Tag en etiquetas definidas para identificación por tipos de palabras encontradas (localización, fecha, persona, entre otros).
- Annotation Set Transfer: Transferencia de anotaciones a grupos específicos de interés.

5.2. Pipeline de análisis

Como se ilustra en la Figura 15 la ejecución del análisis de los textos se realiza mediante la utilización de una aplicación pipeline desarrollada en la herramienta GATE Developer. A continuación, se brinda detalle de los artefactos construidos o implementados para la ejecución de este pipeline:

- Document Reset PR: Es un recurso de procesamiento ANNIE que provee GATE. Es implementado con el propósito de eliminar de un texto anotaciones generadas a partir de otros procesos o anotaciones propias del documento. Para este trabajo se mantiene su configuración por defecto manteniendo anotaciones originales del documento (parámetro: keepOriginalMarkupAS) y sin especificar anotaciones a remover (parámetro: annotationTypes).
- ANNIE English Tokeniser: Es un recurso de procesamiento ANNIE que provee GATE. Es implementado con el propósito de segmentar el texto en pequeñas fracciones o tokens, cada token puede ser palabra, que se identifica con: tipo, texto, longitud y si es una palabra que inicia en letra capital; número, que se identifica con: tipo, número y longitud; y signos de puntuación, que se identifican con: tipo, símbolo y longitud. Este recurso no tiene parámetros de configuración y es de uso obligatorio si se requiere emplear English POS Tagger.
- historias_gazetteer: Es la implementación del recurso de procesamiento ANNIE Gazetteer. Tiene como propósito realizar reconocimiento de entidades a partir de las listas de términos definidas para el procesamiento del texto.

  Algunas de las listas creadas para este trabajo son de uso común, como ciudades, departamentos, números, unidades de tiempo, entre otras. Listas de este tipo pueden ser compartidas entre diferentes proyectos del mismo tipo. Sin embargo, la particularidad en este caso es brindar un método para extraer información a partir de categorías definidas en el documento del Centro Nacional de Memoria Histórica. Por esta razón se han creado las listas descritas en tabla 10.

Tabla 10 Listas destacadas para el proyecto

Nombre	Tipo mayor	Tipo menor	Descripción
_asesinato-Selectivo.lst	claseVictima	_1_4_1_asesinatoSelectivo	Listas definidas para determinar categorías dentro de las que se clasifica el texto.
_masacre.lst	claseVictima	_1_4_2_masacre	
_seviciaTor-tura.lst	claseVictima	_1_4_3_seviciaTortura	
_victima-rios.lst	organization	victimarios	Utilizada para identificar nombres particulares de victimarios, es posible incluir seudónimos si de interés para quien analiza resultados.

Un ejemplo de la definición de las palabras asociadas a las listas se observa en la Figura 14. Particularmente para las listas generadas a partir de la información del documento "¡Basta ya!" [17] se realizó la asignación de pesos a cada una de las palabras, con el objetivo de que estos pesos se tomaran como base para la supervisión y clasificación de las historias.

Dentro del alcance del prototipo se definieron tres categorías, cada una de ellas con igual estructura [Palabra; Peso], como se ilustra en la Figura 16. Para este caso las palabras y los pesos fueron definidos sin acompañamiento de un experto, como parte del ejercicio de generar un modelo para la clasificación. Como resultado, se define que para determinar las clases asociadas a la historia es necesario que un experto defina un diccionario por categoría de interés y que cada término del diccionario tenga un peso definido, el cual podrá ser modificado en el momento de la clasificación como parte de la supervisión realizada a los resultados arrojados por el software prototipo desarrollado.

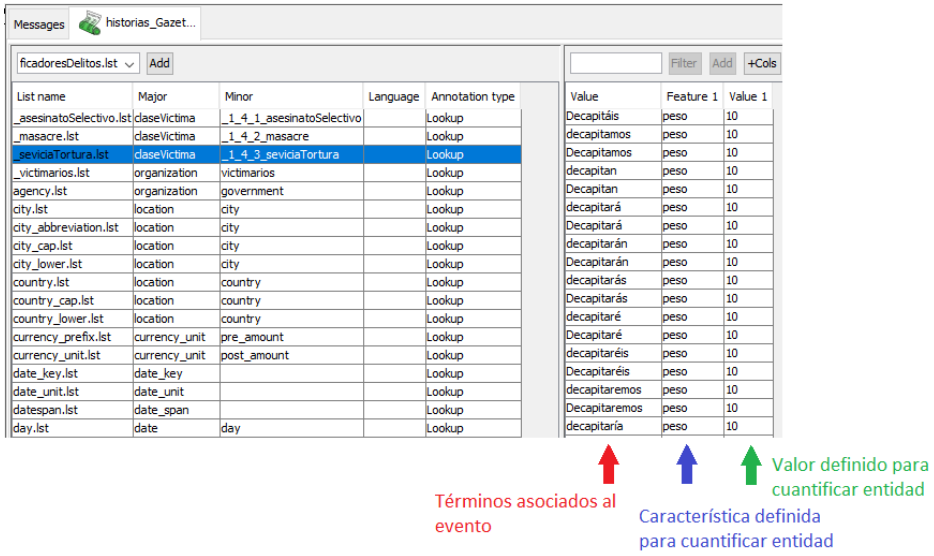


Figura 16 Gazetteer Categoría Sevicia y Tortura

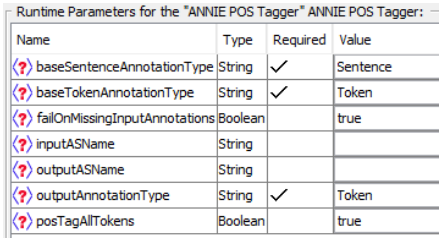
- historias_Jape: Reglas de asociación implementadas con el propósito de extraer las entidades de interés. Para el caso de las categorías de clasificación, se creó un archivo en lenguaje Jape para cada una de ellas. El fin era trasladar a grupos propios de este trabajo cada uno de los fragmentos de texto que se identifican a partir del reconocimiento de entidades. En cuanto a las categorías de clasificación, se entiende que hay unos fragmentos de texto dentro de los que se encuentra la entidad reconocida y un peso asociado a la misma; estos elementos generan un porcentaje de inclusión de la historia completa dentro de las categorías definidas.

  Un ejemplo de esta implementación se muestra en la Figura 17, en donde se observa que a partir de la entidad reconocida (Reglas de lado izquierdo - LHS) se plantea crear una nueva categoría con el respectivo peso (Reglas de lado derecho - RHS).

```
9  Phase:  clasificadores_1_4_3
10 Input: Token Lookup
11 Options: control = appelt
12
13 Rule: clase
14                                                    LHS
15 (
16 {Lookup.minorType == _1_4_3_seviciaTortura}
17 ):clase
18 -->                                                          RHS
19 :clase
20  {
21     gate.AnnotationSet seviciaTorturaSet= (gate.AnnotationSet)bindings.get("clase");
22     gate.Annotation seviciaTorturaFeature = (gate.Annotation)seviciaTorturaSet.iterator().next();
23     gate.FeatureMap newFeatures= Factory.newFeatureMap();
24     newFeatures.put("rule","clase");
25     newFeatures.put("peso",seviciaTorturaFeature.getFeatures().get("peso"));
26     //Crea una nueva etiqueta llamada entidad sevicia tortura
27     outputAS.add(seviciaTorturaSet.firstNode(), seviciaTorturaSet.lastNode(), "entidad_sevicia_tortura", newFeatures);
28  }
```

Figura 17 Implementación reglas Jape

- ANNIE Sentence splitter: Recurso de procesamiento implementado porque es utilizado por los recursos tagger. Se implementa sin valores particulares dentro de sus parámetros de configuración, ya que estos no son requeridos.
- Tagger Number Spanish. Recurso tagger para el etiquetamiento de números. Es un plugin que se utiliza con la intención de diferenciar, mediante reglas de asociación, las entidades de masacre y asesinato selectivo, las cuales se distinguen por el número de muertes que se señalan dentro del texto. Sin embargo, la regla de asociación no fue implementada.
- ANNIE POS Tagger: Recurso tagger para la implementación de etiquetamiento de la oración. No se encontró un recurso bien definido dentro de GATE para realizar esta labor en idioma español. Finalmente se utiliza y se da alcance para identificar localizaciones, las cuales son reconocidas a partir de los Gazetteer de ciudades o departamentos. Como se ilustra en la figura 18 no se especifican valores a parámetros de configuración diferentes a los que se definen por defecto.


Runtime Parameters for the "ANNIE POS Tagger" ANNIE POS Tagger:

Name	Type	Required	Value
baseSentenceAnnotationType	String	✓	Sentence
baseTokenAnnotationType	String	✓	Token
failOnMissingInputAnnotations	Boolean		true
inputASName	String		
outputASName	String		
outputAnnotationType	String	✓	Token
posTagAllTokens	Boolean		true

Figura 18 Parámetros de configuración de ANNIE POS Tagger

Dentro del diseño del pipeline se identificó que este recurso es muy preciso si el texto se encuentra en inglés, lo cual no resulta útil dado el planteamiento de clasificación supervisada de historias propias de Colombia y que su utilidad se desvirtúa cuando se realiza análisis orientado al conocimiento.

- ANNIE NE Transducer: En consecuencia, con la utilidad brindada a *ANNIE POS Tagger*, es implementado con el propósito de identificar la localización. Está dentro de la funcionalidad de este recurso ANNIE realizar identificación de localización según su tipo, el cual podría ser: *region*, *airport*, *city*, *country*, *province*, entre otros, definidos a partir de los Gazetteer implementados. Para este trabajo únicamente se da alcance a determinar qué entidades son de localización sin distinción de tipo.
- Annotation Set Transfer: Con el fin de dar cumplimiento a los objetivos de este trabajo, resultan dos categorías dentro de las cuales se transfieren los resultados de interés. Se transfieren resultados de categorías definidas como tipos de violencia y resultados de elementos comunes como actores y localización. Por esta razón se implementan dos recursos Annotation Set Transfer, que solo reciben como configuración las anotaciones de interés a trasladar, como entrada (annotationTypes), y el nombre de la agrupación de interés, como salida (outputASName). Los grupos de interés para este trabajo son definidos como “clase historias” (asesinato selectivo, masacre o sevicia y tortura) y “comunes” (localización y actores).

6. CONCLUSIONES

- En cuanto a utilidad el trabajo es representativo para entidades del estado, pues una clasificación orientada a categorías apoya el proceso de análisis imparcial de la información, prometiendo generar informes en menor tiempo de lo que actualmente toma.
- El método mediante el cual se definen las categorías y el requerimiento de asignar pesos a las palabras en diccionarios es aceptado por los evaluadores, evidenciado esto en las entrevistas ejecutadas como parte de la etapa de validación.
- En comparación con otras herramientas en el mercado, el trabajo tiene ventajas por ser de licencia pública y por permitir definir categorías verdaderamente relevantes para el contexto.
- Los modelos "Design Science Research Model" y "CRISP-DM" fueron apropiadamente integrados para este trabajo. Siguiendo lineamientos de DSRM se entrega un artefacto de software y se brindan detalles del proceso de construcción en los formatos establecidos por CRISP-DM.
- La relevancia del trabajo es destacada, evaluadores y comunidad académica lo confirmaron en diferentes momentos de la ejecución.
- Se cumplen los objetivos, aunque el trabajo presenta debilidad en profundidad técnica. Como trabajos futuros es recomendado profundizar según las siguientes metas:

Tabla 11 Metas para trabajos futuros

Meta	Descripción
Formas de visualización de la información encontrada dentro del texto.	En las entrevistas con evaluadores se determinó que presentar resultados a partir de porcentajes de inclusión de la historia dentro de un conjunto de categorías, no era suficiente para que un experto realizara con facilidad análisis del contenido. Por lo tanto, es deseable tener gráficas u otros elementos de presentación de resultados que permitan identificar, entre otras, las tendencias que ha tenido el conflicto.
Elementos de control.	Es necesario mejorar los resultados del artefacto, de tal manera que sea posible tener control tanto del texto que se evidencia, a partir del reconocimiento de entidades, así como del que este mismo reconocimiento oculta o deja de encontrar por falta de información dentro del contenido.
Profundizar en técnicas de preparación de textos.	El trabajo se queda corto en resultados técnicos porque ha llegado únicamente a reconocer entidades y aplicar reglas de asociación a partir de patrones. De antemano se conoce que se puede realizar análisis de sentimientos o identificación de actores aplicando técnicas más precisas.

Términos ambiguos	Dentro del prototipo final no se han aplicado técnicas para identificar términos que son ambiguos dentro de las categorías, por lo cual es importante recrear elementos de control que puedan mejorar estos resultados.

- El análisis de texto o información no estructurada es importante para mantener los detalles del contenido. El trabajo se destaca por aprovechar la información no estructurada sin proponer un cambio al respecto, lo cual beneficia la labor que realizan aquellos expertos que se involucran directamente con las víctimas, ya que ahorran tiempo valioso si no se ocupan de los detalles que devenga registrar los resultados en un sistema de información completamente estructurado.
- El potencial y la innovación de este trabajo son destacados y es importante afianzar la relación con Defensoría del Pueblo y Unidad de víctimas. Ellos en su calidad de expertos en el tema brindan desde los requerimientos unos objetivos importantes para enriquecer el trabajo.

7. REFERENCIAS

[1] "Defensoría Delegada para la Orientación y Asesoría de las Víctimas del Conflicto Armado interno," *Defensoría del Pueblo*. [Online]. Available: http://defensoria.gov.co/es/public/defensoriasdelegadas/1448/Defensoría-Delegada-para-la-Orientación-y-Asesoría-de-las-Víctimas-del-Conflicto-Armado-interno.htm. [Accessed: 15-Nov-2016].

[2] "Peace Technology: Scope, Scale, and Cautions," *Building Peace*. .

[3] "Joining the Storytelling Revolution: Interview with International Storytelling Center President Kiran Singh Sirah," *Building Peace*. .

[4] J. J. Bocanegra García, R. A. González, and L. Olaya Bello, "Una estrategia para la apropiación de las TIC en la reconciliación de las víctimas del conflicto armado colombiano," *TRILOGÍA Cienc. Tecnol. Soc.*, vol. 8, no. 14, pp. 53–64, Jan. 2016.

[5] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to Analyze Political Attention with Minimal Assumptions and Costs," *Am. J. Polit. Sci.*, vol. 54, no. 1, pp. 209–228, Jan. 2010.

[6] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Polit. Anal.*, vol. 21, no. 3, pp. 267–297, Jul. 2013.

[7] D. M. Harikrishna and K. S. Rao, "Children story classification based on structure of the story," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 1485–1490.

[8] D. Nguyen, D. Trieschnigg, and M. Theune, "Folktale Classification Using Learning to Rank," in *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds. Springer Berlin Heidelberg, 2013, pp. 195–206.

[9] P. Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization," *Int. J. Mach. Learn. Comput.*, pp. 423–426, 2012.

[10] D. Beeferman, A. Berger, and J. Lafferty, "Statistical Models for Text Segmentation," *Mach. Learn.*, vol. 34, no. 1–3, pp. 177–210, Feb. 1999.

[11] J. M. Ponte and W. B. Croft, "Text segmentation by topic," in *Research and Advanced Technology for Digital Libraries*, C. Peters and C. Thanos, Eds. Springer Berlin Heidelberg, 1997, pp. 113–125.

[12] G. Wiedemann, "Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences," *Hist. Soc. Res. Hist. Sozialforschung*, vol. 38, no. 4 (146), pp. 332–357, 2013.

[13] A. P. Rafael A. Gonzalez, "La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería," 2012.

[14] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[16] J. C. M. Zapata and N. Gil, "Incorporation of both pre-conceptual schemas and goal diagrams in CRISP-DM," in *Computing Congress (CCC), 2011 6th Colombian*, 2011, pp. 1–6.

[17] "¡Basta ya! Colombia: memorias de guerra y dignidad," *¡Basta ya! Colombia: memorias de guerra y dignidad*. [Online]. Available: http://www.centrodememoriahistorica.gov.co/micrositios/informeGeneral/descargas.html. [Accessed: 02-Sep-2016].

[18] K. Crowston, E. E. Allen, and R. Heckman, "Using natural language processing technology for qualitative data analysis," *Int. J. Soc. Res. Methodol.*, vol. 15, no. 6, pp. 523–543, 2012.

[19] R. A. Gonzalez and S. Henk G., "'Validation and design science research in information systems.' Research methodologies, innovations, and philosophies in software systems engineering and information systems.," in *Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems*, IGI Global Information Science Reference, 2012, pp. 403–426.

[20] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer-Assisted Text Analysis for Comparative Politics," *Polit. Anal.*, vol. 23, no. 2, pp. 254–277, Apr. 2015.

[21] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, I. S. Vizirianakis, and I. Iliopoulos, "DrugQuest - a text mining workflow for drug association discovery," *BMC Bioinformatics*, vol. 17, no. 5, p. 182, Jun. 2016.

[22] H. BORJA OROZCO, I. BARRETO, J. M. SABUCEDO, and W. LÓPEZ, "Construcción del discurso deslegitimador del adversario: gobierno y paramilitarismo en Colombia." [Online]. Available: http://www.scielo.org.co/pdf/rups/v7n2/v7n2a20.pdf. [Accessed: 21-Aug-2016].

[23] R. Harpaz *et al.*, "Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art," *Drug Saf.*, vol. 37, no. 10, pp. 777–790, Oct. 2014.

[24] I. Korkontzelos, D. Piliouras, A. W. Dowsey, and S. Ananiadou, "Boosting drug named entity recognition using an aggregate classifier," *Artif. Intell. Med.*, vol. 65, no. 2, pp. 145–153, Oct. 2015.

[25] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics," *PLoS Comput. Biol.*, vol. 9, no. 2, Feb. 2013.

[26] H. Cunningham, D. Maynard, and K. Bontcheva, *Text Processing with GATE*. Gateway Press CA, 2011.

[27] A. Dehghan, A. Kovacevic, G. Karystianis, J. A. Keane, and G. Nenadic, "Combining knowledge- and data-driven methods for de-identification of clinical narratives," *J. Biomed. Inform.*, vol. 58, Supplement, pp. S53–S59, Dec. 2015.