Aumento de datos para tareas relacionadas al perfilado de autor

Tesis de Maestría

Por:

Victor Jimenez Villar

Asesor:

Dr. Luis Villaseñor Pineda

Instituto Nacional de Astrofísica Óptica y Electrónica
Coordinación de Ciencias Computacionales

Tonantzintla, Puebla. Julio 2019

Agradecimientos

Esta investigación fue realizada gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), a través de la beca No. 868585.

Dedicatoria

A mi familia y amigos, por motivarme cada día superar mis límites.

Resumen

Abstract

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1
Introducción

Imagina que se te ha dado un texto de un autor anónimo, y deseas saber tanto como sea posible del autor (género, ocupación, personalidad etc.), solo analizando el texto dado. El interes en perfilado de autor ha ido creciendo gracias al constante flujo de informacion compartida atraves de redes sociales (por ejemplo Twitter, Facebook, Reddit) y su aplicaciones varian desde mercadotecnia hasta seguridad nacional. Existen numerosas razones del porque nos interesa conocer datos relevantes de los usuarios de redes sociales. Por ejemplo, a las empresas les interesaria conocer a que tipo de usuarios les gusta o no su producto o servicio, con la intencion de dirigir una mejor campaña de publicidad Ikeda et al. (2013). Ademas en un contexto de seguridad informatica a la policia cibernetica le gustaria conocer el perfil de la personas que envian mensajes amenzantes o de acoso sexual Bogdanova, Rosso, y Solorio (2012).

Diversos estudios dentro de la comunidad socioliguistica han demostrado que las palabras que las personas utilizan en sus vidas diarias pueden revelar importantes aspectos sociales y psicologicos. Con avances en computación, el analisis de textos permite a los investigadores a obtener caracteristicas de lo que las personas dicen y tambien de las particularidades en sus estilos linguisticos Pennebaker, Mehl, y Niederhoffer (2002).

Debido al lenguaje informal de redes sociales y poco estandarizado hace que esta tarea sea desafiante, por ejemplo: errores gramaticales, abreviaturas, anglicismos, emoticones o incluso texto generado por cuentas automáticas. Una de las conferencias más destacadas en perfilado de autor ha sido el PAN@CLEF (una serie de eventos científicos y tareas compartidas en el análisis forense digital y estilo métrico); desde el año 2013 al actual 2020 se han estudiado diversos enfoques del perfilado de autor desde una perspectiva multi-idioma (inglés y español principalmente) entre las cuales destacan: Identificación de edad y género Rangel et al. (2013), identificación

de personalidad, variación de lenguaje y dimensión de género Stammatatos et al. (2015). Ademas de estas tareas en las conferencias ERISK se han investigado tema mas complejos como lo son la identificacion de depresión y anorexia Losada, Crestani, y Parapar (2018), aunque no se menciona explicitamente como perfilado de autor, los rasgos psicologicos y estado emocional son caracteristicas importantes dentro del perfilado.

1.1. Planteamiento del problema

Para resolver las tareas de perfilado de autor, la mayoria de los trabajos existentes se han enfocado en utilizar algoritmos de aprendizaje computacional en convinación con diferentes tecnicas para extraer caracteristicas: conteo de palabras Laserna, Seih, y Pennebaker (2014), identificación de frases personales Ortega-Mendoza et al. (2018), analisis de emociones Aragón et al. (2019) entre otras tecnicas. La obtención de dichas caracteristicas requiere un analisis riguroso y en muchas ocaciones es necesaria la intervención de expertos en el tema. Sin embargo existen tecnicas de apredizaje computacional mas complejas como las redes neuronales en donde la extracción de caracteristicas se realiza de forma automatica mediante una serie de abstracciones.

La principal motivación para el uso de redes neuronales en perfilado de autor, es debido al increible exito del aprendizaje profundo en tareas complejas para el entendimiento del lenguaje : parafraseo, traducción automatica, analogia, implicació textual, similitud semantica, etc. En el conjunto de datos GLUE **?** los modelos de aprendizaje pronfundo han superado la puntuación humana, Christopher D. Manning menciona que desde el año 2015 se produjo un tsunami de deep learning en el area procesamineto de lenguaje natural, debido a la gran cantidad de papers en coferencias de NLP utilizando aprendizaje profundo Manning (2015).

De acuerdo al estado de arte en la ultima conferencia del PAN@CLEF los equipos con mejores resultados utilizaron tecnicas tradicionales de aprendizaje como lo son maquinas de soporte vectorial SVM en combinación con n-gramas de caracteres. Asi tambien en las tareas del ERISK el mejor performance se obtuvo extrayendo caracteristicas en convinacion con un ensamble de bolsas de palabras BOW y diferentes clasificadores. Lo que se ha podido observar en los diferentes reportes de estas conferencias, es que los modelos de aprendizaje basados en redes neuronales no han tenido el exito esperado.

Uno de los principales problemas dentro del campo de aprendizaje automatico es que el exito de este depende de la cantidad de datos etiquetados con que se cuente y se hace mas notable cuando se utilzan modelos de aprendizaje profundo, el etiquetado manual de datos consume mucho tiempo y es costoso, ademas se podria comprometer a problemas legales debido al uso de datos personales como es el caso en las tareas de perfilado de autor, los estudios en la literatura tratan con un numero pequeño de autores conocidos, en donde el eitquetado manual puede ser aplicado, pero considerando las dimensiones de los datos en redes sociales se convierte en una tarea mas dificil.

Otro problema conocido y estudiado es el sobreajuste en la etapa de entrenamiento significando una gran diferencia entre los resultados del modelo en entrenamiento y los resultados en el conjunto de prueba, para esto se han propuesto diferentes tecnicas como lo son el dropout o agregar ruido aleatorio a los ejemplos originales.

Observando las limitantes anteriores este trabajo presenta un estudio sobre el efecto de agregar mas documentos sinteticos, mediante aumento de datos, al conjunto de entrenamiento original y el efecto que tiene en los algoritmos de redes neuronales aplicados en tareas relacionadas al perfilado de autor. El aumento de datos es una via para obtener mas documentos de forma automatica y tambien util en la regularización de los modelos de aprendizaje pronfundo.

Algunas de las principales preguntas a contestar son :

1.- ¿Que proporción de modificación a nivel parrafo (post o tweet) es el ideal?

2.- ¿Que efecto tiene realizar diferentes proporciones de modificación?

3.- ¿Es conveniente agregar pocos documentos con mucha modificación o varios documentos con poca modificación?

4.- ¿Que efecto tiene el aumento de datos en un contexto considerando el contexto global del texto?

5.- ¿Que efecto tiene el aumento de datos en un contexto considerando un contexto local en el texto?

6.- ¿El aumento de datos permite reducir el sobre ajuste en tareas relacionadas al perfilado de autor?

•

1.2. Objetivos

En este proyecto de tesis se plantean los siguientes objetivos.

1.2.1. Objetivo general

Proponer un método de aumento de datos, cuando se cuenta con pocos datos etiquetados, para mejorar la predicción de los modelos de aprendizaje profundo en las tareas de perfilado de autor.

1.2.2. Objetivos específicos

1.- Determinar una forma confiable y diversa de realizar aumento de datos para tareas de perfilado de autor.

2.- Determinar un modelo apropiado de aprendizaje profundo que aproveche el aumento de datos.

3.- Analizar el impacto del aumento de datos para perfilado de autor.

1.3. Organización de la tesis

Esta tesis esta organizada de la siguiente forma:

- Capitulo 2: **Marco teorico**; Presenta una rapida introduccion a la clasificación de textos con aprendizaje automatico, ademas de mencionar las principales metricas de evaluación utilizadas en este trabajo. Los conceptos descritos son fundamentales para comprender la solución propuesta.
- Capitulo 3: **Trabajo relacionado**; Describe el estado del arte en perfilado de autor y aumento de datos para clasificación de textos, su principal objetivo es conocer como se ha abordado el problema ademas de analizar sus pros y contras.

Capítulo 2

Marco Teorico

En este capitulo se describen conceptos relacionados a la tarea de perfilado de autor mediante algoritmos de aprendizaje automatico. Se describen las principales representaciones de un texto dado, las caracteristicas generales de los clasificadores empleados, asi como las medidas de evaluacion empleadas para medir los resultados de los diferentes modelos. Ademas se presenta una introducción de las tareas de procesamiento de lenguaje natural utilizadas en el metodo propuesto: Etiquetado de partes de la oración, parafrasis y resolución de analogias.

2.1. Clasificación de textos

En años recientes, ha habido un crecimiento exponecial en el numero de documentos complejos y textos que no se pueden procesar por medios manuales, tal es el caso del perfilado de autor, en donde se desea conocer la categoria (clase o grupo de autores) a la que pertenece un documento dado (historial del usuario). Los problemas de clasificación de textos han sido ampliamente estudiados en las ultimas decadas, expecialmente con los recientes avances en procesamiento de lenguaje natural, muchos investigadores estan interesados desarrollar aplicaciones que mejoren los metodos de clasificación de textos.

La clasificación de textos puede describirse en cuatro pasos: extracción de caracteristicas, selección de caracteristicas, seleccion del clasificador y evaluación.

2.2. Extracción de caracteristicas

El pre-prpocesamiento y la extracción de caracteristicas son pasos muy importantes en la classificación de textos, en las siguientes secciones se presentan algunas de las tecnicas mas empleadas y se mencionan dos metodos de representación de caracteristicas el pesado de palabras y los vectores de palabras.

2.2.1. Pre-procesamiento

La mayoria del texto existente contiene palabras inecesarias, para algunas tareas de clasificación como lo son: palabras de paro, errores gramaticales, signos de puntuación etc. Ademas de esto el texto extraido de redes sociales cotiene enlaces de internet, menciones de usuario, etiquetas (conocidos como hasgtags), emoticones y un vocabulario muy informal. A continuación se explica brevemente algunas tecnicas empleadas para el limpiado y pre-procesamiento de textos.

1.- **Tokenización**: Es un metodo de pre-procesamiento en el cual se divide una cadena de caracteres en palabras , frases, simbolos y otros elementos dentro del texto llamdos tokens. Se pueden utilizar diferentes algoritmos para poder realizarlo lo mas simple es separar el texto mediante un espacio o caracter comun, por ejemplo:

   Texto original: *"Los dias de verano son calurosos"*.

   Los tokens del texto anterior son los siguientes: {"Los", "dias", "de", "verano", "son", "calurosos"}'

2.- **Palabras de paro**: Son palabras con mayor frecuencia en los documentos, tales como: {"a" "the", "they", "he" , "she", ...} (Para el idioma Inglés). En algunas tareas de clasificación de textos las palabras de paro no son de importancia y lo mas comun es removerlas de los documentos o textos.

3.- **Capitalización**: Dado que los documentos consisten en muchas oraciones, existe una capitalización de palabras diversa, lo mas coumun es reduccir todas las letras a minusculas.

4.- **Reducción de ruido**: La mayorias de los textos contienen muchos caracteres innecesarios, como signos de puntuación o caracteres especiales. En tareas como

detección de autoria pueden ser utiles pero en muchas ocaciones solo agregan ruido a los modelos de clasificación de textos.

5.- **Otras tecnicas**: Adicionalmente a las tecnicas descritas se encuentran; corrección de errores ortograficos, lematización y stemming.

2.2.2. Bolsa de palabras BoW

El modelo de bolsa de palabras o BoW (por sus siglas en ingles "Bag of Words"), es una versión reducida y simplificada de un texto, basado en un criterio especifico, como lo puede ser mediante la frecuencia de cada palabra.

En el modelo BoW, el conjunto de documentos es representado mediante una matriz de numeros, siendo las columnas palabras unicas del conjunto de datos y las filas cada documento. Las palabras no se representan en forma secuencial, como en una oración o un documento, y las relaciones semanticas entre las palabras se pierden. En este modelo las palabras representan el contenido de un documento y pueden ser utilizadas para determinar su tema principal.

Ejemplo de BoW

Texto original: " *Informalmente, un algoritmo es cualquier procedimiento computacional bien definido que recibe algun valor, o conjunto de valores, como entrada y produce un valor, o conjunto de valores, como salida. Un algoritmo es entonces un conjunto de pasos computaciones que transforman la entrada en la salida.*".

Bolsa de palabras: { procedimiento, valores, algoritmo, bien, de, o, que, algun, como, valor, salida, transforman, pasos, recibe, computaciones, computacional, ,, Un, la, definido, cualquier, en, Informalmente, es, produce, ., un, entonces, entrada, conjunto, y}

Representación de caracteristicas: [1, 2, 2, 1, 3, 2, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 5, 1, 2, 1, 1, 1, 1, 2, 1, 2, 3, 1, 2, 3, 1]

Pesado de palabras

La forma mas basica de extracción de caracteristicas es mediante el pesado TF, el cual consiste en contar el numero de ocurrencias de cada palabra en el conjunto de datos completo. Los metodos basados en TF generalmente consisten en representar la frecuencia de palabras como un peso escalado o normalizado, aunque es facil imple-

metación y muy intutivo este metodo esta limitado por el hecho de que las palabras mas comunes pueden dominar la representación.

TF-IDF Term Frequency-Inverse Document Frequency

Esta tecnica de pesado fue propuesta por K. Sparck Jones **?**, con el objetivo de mitigar el efecto de las palabras mas comunes en el corpus. IDF asigna menos peso a palabras con alta frecuencia en toda la colección de documentos. La representación matematica del peso de un termino en un documento por TF-IDF esta dada en la ecuacion **??**

eq. W(d,t)= TF(d,t)*log (N/df(t))

En donde N es el numero de documentos y $df(t)$

2.2.3. Vectores de palabras

2.2.4. Glove

2.2.5. FastText

2.3. Selección de caracteristicas

2.3.1. Xi Cuadrada

2.4. Selección del clasificador

INTRO

2.4.1. Maquina de Vectores de Soporte (SVM)

2.4.2. Aprendizaje profundo

2.4.3. Redes Neuronales Profundas (DNN)

2.4.4. Redes Neuronales Reccurrentes (RNN)

2.4.5. Redes Neuronales Convolucionales (CNN)

2.4.6. Limitaciones del aprendizaje profundo

2.5. Metricas de evaluación

Capítulo 3

Trabajo Relacionado

Descubrir las caracteristicas de un autor anonima es de interes para la comunidad cientifica en procesamiento de lenguaje natural. Existen numerosas razones una de ellas es aprovechar el constante flujo de información en redes sociales para entender el mejor el lenguaje coloquial de uso diario, hacer que nuestras maquinas puedan identificar emociones, estados de animo, el genero y edad de una persona , etc.

Muchos esfuerzos y avances se han realizado en la ultima decada, por ejemplo los foros de evaluación PAN@CLEF y ERISK, en este capitulo se explora el trabajo previo en esta area. Dado que no existen estudios sobre perfilado de autor y el aumento de datos, se presenta una revisión general del aumento de datos en tareas de clasificación de textos.

3.1. Perfilado de autor

3.1.1. Detección de depresión y anorexia

3.1.2. Enfoques tradicionales

3.1.3. Enfoques de aprendizaje profundo

3.2. Aumento de datos

3.2.1. Aumento de datos en clasificación de textos

3.2.2. Desventajas del aumento de datos

Capítulo 4

Colecciones de datos y configuración experimental

Capítulo 5

Metodo Propuesto

Capítulo 6

Experimentos y Resultados

Capítulo 7

Conclusiones y Trabajo Futuro

Apéndices

Apéndice

Detalles de los experimentos

Bibliografía

Aragón, M. E.; López-Monroy, A. P.; González-Gurrola, L. C.; y Montes-y Gómez, M. 2019. Detecting Depression in Social Media using Fine-Grained Emotions. (2013):1481–1486.

Bogdanova, D.; Rosso, P.; y Solorio, T. 2012. On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, 110–118. Association for Computational Linguistics.

Ikeda, K.; Hattori, G.; Ono, C.; Asoh, H.; y Higashino, T. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems* 51:35–47.

Laserna, C. M.; Seih, Y. T.; y Pennebaker, J. W. 2014. Um. . Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology* 33(3):328–338.

Losada, D. E.; Crestani, F.; y Parapar, J. 2018. Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). *CEUR Workshop Proceedings* 2125.

Manning, C. D. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics* 41(4).

Ortega-Mendoza, R. M.; López-Monroy, A. P.; Franco-Arcega, A.; y Montes-y Gómez, M. 2018. Emphasizing personal information for Author Profiling: New approaches for term selection and weighting. *Knowledge-Based Systems* 145:169–181.

Pennebaker, J. W.; Mehl, M. R.; y Niederhoffer, K. G. 2002. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology* 54(1):547–577.

Rangel, F.; Rosso, P.; Koppel, M.; Stamatatos, E.; y Inches, G. 2013. Overview of the author profiling task at PAN 2013. *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* 352–365.

Stammatatos, E.; Daelemans, W.; Verhoeven, B.; Juola, P.; López-López, A.; Potthast, M.; y Stein, B. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings* 1391(31):898–927.