Un Método para Establecer las Variables Demográficas de los Usuarios en Tweeter

Ricardo Alexander Moncayo Martínez

13 de Mayo 2022

1 Introducción

El uso del lenguaje revela aspectos intrínsecos de la personalidad, su entorno social, edad[1], afiliación política, gustos, localización geográfica, este se modifica con el tiempo como cualquier otro aspecto de los comportamientos sociales [2, 3]. Las relaciones del uso del lenguaje y aspectos demográficos han sido estudiadas desde hace mucho tiempo[1], lograr establecer estas relaciones es una tarea difícil dada la complejidad de las expresiones humanas, las modificaciones del lenguaje de acuerdo al contexto[4]. Con los últimos avances tecnológicos en el análisis de grandes cantidades de datos y el uso masivo de las redes sociales como twitter y facebook se pueden explorar estas relaciones demográficas [2, 5].

Ests propuesta se enfoca en tres variables demográficas: rango de edad, sexo, ubicación geográfica. Las variables de edad y sexo están entrelazadas, por lo que para lograr establecerlas se deben estudiar simultáneamente, además las personas con el uso del lenguaje eligen o no mostrar su sexo dependiendo de la cultura y el escenario [2], esta variabilidad que pueden tener los diferentes mensajes de una persona hace que construir un modelo de lenguaje sea una tarea laboriosa. [2, 5]. En el trabajo de [4] construyen una base de datos de 2439 usuarios, y anotan manualmente su genero y edad, a partir de los textos de los usuarios realizan una extracción de características basadas en las frecuencias de las palabras más usadas entre las clases y entrenan un modelo de regresión lineal a partir de estos, reportan un F-score de 0.75 para hombres y 0.77 para mujeres. En [6] los autores abordan el problema de la edad dividiendo la población en dos generaciones nacidos en 1979 y 1984, sobre una base de datos de Blogs, se extraen características de comportamiento en linea, estilo del léxico, y contenido de léxico, haciendo uso de una bolsa de palabras construida con las palabras más frecuentes, reportan una precisión entre el $79.95 - 81.5$ un trabajo similar se encuentra en [7] para definir el genero de la persona usando un clasificador. en [8] evalúan dos maquinas de soporte vectorial (SVM) y una red neuronal Bert en una base de datos con múltiples tweets de cada usuario, logrando un f-score de 0.855, una revisión extensa se encuentra en [9].

Ubicación geográfica en tweets ha sido estudiada como método para caracterizar poblaciones como en los Estados Unidos de América en [10], se propone construir relaciones entre los nombres, apellidos, la etnia o raza y genero para establecer las conexiones de los usuarios y el espacio que ocupan. Un esfuerzo de diferentes científicos de datos para desarrollar un modelo que ubique la ubicación geográfica del usuario de tweeter se presento en este reto [11], el primer lugar logro un accuaracy de 0.45 en esta tarea, para esto uso una tokenizacion n-gram, para el texto de los tweets, en estos ubicaron mensajes donde el usuario declara su ubicación para fortalecer el modelo de bolsa de palabras, estas características concatenadas junto con la zona horaria y etiquetas de localización son usadas en un SVM para esta tarea, un enfoque similar a este pero usando una red neuronal para establecer estas relaciones es presentada en [12].

2 Problema

El uso del lenguaje se modifica de acuerdo al contexto o entorno del orador, este puede revelar o no aspectos de la personalidad. Este se ve también influenciado por aspectos sociales como el nivel educativo, la personalidad, los gustos, el origen étnico, el genero, la edad, circulo social, por lo que identificar variables demográficas a partir de lenguaje escrito conlleva a analizar y caracterizar una alta variabilidad presente en los textos.

3 Preguntas

¿Cuales son las características de lenguaje que predominan sobre las variables demográficas en estudio? ¿cómo establecer y diferenciar las ambigüedades semánticas de los mensajes?

3.1 Objetivos

Objetivo general:

Establecer un modelo que sea capaz de predecir aspectos demográficos del usuario, tales son la edad, el genero, la ubicación partir del análisis cuantitativo de los diferentes mensajes contenidos en la base de datos.

Objetivos Específicos:

1. Determinar las estructuras y palabras que son relevantes para determinar las variables demográficas
2. Usando las estructuras y palabras determinadas, extraer las características de los textos
3. Generar un espacio de representación para las variables demográficas.
4. Establecer una correlación entre el espacio de características y las variables demográficas
5. Construir un modelo matemático que determine las características demográficas de edad, genero, localización geográfica a partir de mensajes de tweeter.

4 Metodología

Determinar las estructuras y palabras que son relevantes para determinar las variables demográficas

Actividades:

1. Realizar limpieza de datos sin comprometer información
2. Determinar si la cantidad de datos final es suficiente para caracterizar las variables demográficas
3. Representación de las palabras en forma de tokens
4. Encontrar y caracterizar las palabras más usadas y sus relaciones espaciales en el texto

Métodos: Usar y comprobar la lematización, posibles librerias spacy,Stanford CoreNLP or FreeLing, generar un modelo propio.
Determinar las relaciones espaciales y de frecuencias, análisis apartir de bolsa de palabras, k-means,

probabilistic latent semantic analisis, redes neuronales con modulo de atención, grafos, diccionarios supervisados.

Riesgos: En caso de que la cantidad de datos no sea suficiente se debe buscar más datos de fuentes publicas o privadas y realizar las respectivas anotaciones, también se puede emplear textos de otros problemas. En caso de no ser posible se deberá construir un modelo considerando las posibles limitaciones para la clase menos representada.

Generar un espacio de representación para los textos

Actividades:

1. Establecer una métrica entre las estructuras y palabras encontradas con los textos
2. Evaluar el espacio de representación
3. Proyectar los textos al espacio de presentación

Riesgos: la métrica usada no relaciona correctamente las variables y estructuras del texto, se debe plantear otra métrica o revisar la cantidad de datos.

Métodos: Métricas en el espacio de representación como el coseno, cityblock, son usadas en el análisis de textos, clusterizaciones suaves como los procesos de Dirichlet puede ser empleados para generar el espacio de representación.

Establecer las relaciones entre el espacio de características y las variables demográficas

Actividades:

- Usar el espacio de características para construir un modelo matemático de las variables demográficas estudiadas
- Evaluar y Validar el modelo basado en las anotaciones o ground truth

Metodos: Un classificador de redes neuronales, boosting, establecer métricas como el f1-score, precisión, recall para la validación.

Riesgos: El modelo generado no clasifique correctamente, usar otro clasificador, reducción de dimensionalidad del espacio generado, revisar actividades anteriores.

5 Cronograma

Actividad/Semana	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
Realizar limpieza de datos sin comprometer informaci on	■	■	■	■	■							
Determinar si la cantidad de datos final es suficiente para caracterizar las variables demograficas	■	■	■	■	■							
Representacion de las palabras en forma de tokens		■	■	■	■	■						
Encontrar y caracterizar las palabras m as usadas y sus relaciones espaciales en el texto		■	■	■	■	■	■					
Establecer una metrica entre las estructuras y palabras encontradas con los textos					■	■	■	■	■			
Evaluar el espacio de representacion						■	■	■	■	■		
proyectar los textos al espacio de presentación							■	■	■	■	■	
Usar el espacio de caracterısticas para construir un modelo matem atico de las variables demograficas estudiadas							■	■	■	■	■	■
Evaluar y Validar el modelo basado en las anotaciones o ground truth							■	■	■	■	■	■

Bibliografía

[1] W. Labov, *The social stratification of English in New York city*, Cambridge University Press, 2006.

[2] S. E. Wagner, "Age grading in sociolinguistic theory," *Language and Linguistics Compass* **6**(6), pp. 371–382, 2012.

[3] D. Bickerton, "Peter trudgill, the social differentiation of english in norwich.(cambridge studies in linguistics 13.) cambridge: Cambridge university press, 1974. pp. x+ 211.," *Journal of linguistics* **11**(2), pp. 299–308, 1975.

[4] D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong, "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment," in *25th International Conference on Computational Linguistics (COLING 2014)*, pp. 1950–1961, Dublin City University and Association for Computational Linguistics, 2014.

[5] A. E. Marwick and D. Boyd, "I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience," *New media & society* **13**(1), pp. 114–133, 2011.

[6] S. Rosenthal and K. McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 763–772, 2011.

[7] Z. Miller, B. Dickinson, and W. Hu, "Gender prediction on twitter using stream algorithms with n-gram character features," 2012.

[8] A. Z. Klein, A. Magge, and G. Gonzalez-Hernandez, "Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets," *PloS one* **17**(1), p. e0262087, 2022.

[9] S. Sharma and V. Gupta, "Role of twitter user profile features in retweet prediction for big data streams," *Multimedia Tools and Applications* , pp. 1–30, 2022.

[10] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of twitter users," in *Proceedings of the International AAAI Conference on Web and Social Media*, **5**(1), pp. 554–557, 2011.

[11] B. Han, A. Rahimi, L. Derczynski, and T. Baldwin, "Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text," in *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 213–217, The COLING 2016 Organizing Committee, (Osaka, Japan), Dec. 2016.

[12] P. Thomas and L. Hennig, "Twitter geolocation prediction using neural networks," in *Language Technologies for the Challenges of the Digital Age*, G. Rehm and T. Declerck, eds., pp. 248–255, Springer International Publishing, (Cham), 2018.