Midiendo la controversia en redes sociales a través de la jerga

Juan Manuel Ortiz de Zarate[1] and Esteban Feuerstein[1,2]

[1] Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
[2] Fundación Sadosky, Argentina

Abstract. En este trabajo desarrollamos una metodología para cuantificar la *controversia* en una red social exclusivamente a través de su jerga, es decir por el lenguaje que utilizan sus usarios. Presentamos resultados preliminares de nuestros experimentos en los que logramos una accuracy equiparable a los métodos del estado del arte basados en la estructura.

Keywords: PLN - Detección de comunidades - Redes Sociales - Controversia

1 Introducción

La polarización es un fenómeno de alto impacto a distintos niveles, como se observa en diversos casos registrados en la literatura de distintas disciplinas desde hace muchas décadas, desde la división de un club de karate[25] hasta cuestiones políticas[17, 18, 3] o parlamentarias[7]. Con la aparición de las redes sociales mediadas tecnológicamente [9] podemos ver cómo éstas pueden ser intervenidas intencionalmente [21, 6] o cómo pueden derivar en ataque virtuales [16]. Todos estos trabajos marcan distintos problemas aparejados por la polarización y se proponen y/o preguntan cómo resolverlos. Por ejemplo Kumar, Srijan, et al.[16] proponen y analizan distintas estrategias para defenderse de los ataques entre comunidades mientras que Stewart, et al.[21] insinúan que hubo injerencia externa para agudizar la polarización y de esta manera beneficiar a un candidato en particular. El concepto de controversia tiene diversas potenciales aplicaciones en los escenarios informativos y debates públicos. Detectar la controversia puede proveer bases para analizar la *dieta de noticias* de los lectores [15, 19], ofreciendo la chance de mejorar la información proveyendo recomendaciones de contenido a leer[22] y/o conectando a usuarios con puntos de vista opuestos [2, 10]

Los trabajos presentados anteriormente comparten una limitación: todos ellos son casos de estudio específicos, donde la controversia está identificada en un solo conjunto de datos cuidadosamente seleccionados y recolectados a través de un amplio conocimiento del dominio. Con el objetivo de superar estas limitaciones nos proponemos profundizar el trabajo hecho por Garimella et al.[12], donde desarrollan distintas métricas contexto-agnósticas para cuantificar la controversia, tanto desde un punto de vista estructural de la red como desde su contenido. El mencionado artículo presenta muy buenos resultados con enfoques basados en la estructura del grafo pero en cuanto intentan usar el texto de los posteos los resultados no son satisfactorios, pues no logran una métrica que diferencia casos con y sin polarización.

En este trabajo proponemos una nueva metodología para obtener una cuantificación de la controversia utilizando exclusivamente el texto. A través de modelos predictivos de NLP y el cáculo de un score basado en *Dipole Moment*[3] logramos una muy buena diferenciación entre casos con y sin controversia. Para nuestros experimentos utilizamos las mismas discusiones de [12] además de otras más recientes recolectadas por nosotros.

2 Metodología

Nuestro enfoque está basado en un forma sistemática de caracterizar la actividad en las redes sociales. Para esto definimos un *pipeline* de 4 etapas: construcción del grafo, identificación de comunidades, construcción del modelo predictivo y cuantificación de la controversia, donde el resultado final es un valor que mide cuan controversial es la discusión.

Al igual que en [12] utilizamos Twitter para nuestros experimentos dada la gran disponibilidad de información que esta plataforma nos provee. A través de su API[3] descargamos un conjunto de tweets relacionados a un determinado hashtag o *key word* de la discusión. En la primer etapa construimos un *grafo direccionado* donde cada nodo es un usuario que *twiteó* en la conversación y las aristas representan *retweets* entre ellos apuntando hacia el usuario que fue *retweeteado*. Mediante este criterio de selección de los ejes nos proponemos representar la idea de *endorsment* entre los actores.

En el segundo paso utilizamos *Walktrap*[20] para identificar las distintas comunidades. Este método resulta muy efectivo desde un punto de vista práctico, además de funcionar bien de acuerdo a medidas generalmente aceptadas como la modularidad Q[4]. Intuitivamente en nuestro caso, dada la semántica de los ejes, los usuarios serán particionados en comunidades, conjuntos disjuntos de nodos con posiciones distintas en el debate.

En el tercer paso tomamos todos los tweets emitidos por los usuarios de las dos comunidades más grandes devueltas por *Walktrap* y los normalizamos [5]. Posteriormente agrupamos todos los tweets de un mismo usuario en una sola cadena de texto y le asignamos la etiqueta *C1* o *C2* basándonos en la comunidad de pertenencia. Finalmente utilizamos estos datos como set de entrenamiento de nuestro modelo de *NLP*. Para la construcción de este modelo utilizamos el algoritmo denominado *FastText*[6] que se destaca por equiparar el *accuracy* de los más modernos modelos de Deep Learning[1, 24, 23, 4] con un costo computacional notablemente mejor.

En el cuarto y último paso utilizamos el clasificador para predecir la clase de cada usuario de la principal componente conexa. Para esto, concatenamos todos los tweets de un mismo usuario en una sola cadena de texto. Teniendo en cuenta que *FastText* nos brinda una cuantificación de la predicción, es decir una medida de cuán certera cree que es la predicción, nos quedamos con el subconjunto de usuarios cuya pertenencia a una comunidad tiene probabilidad mayor de 90%.

Finalmente para establecer el *score de controversia* definimos la métrica a la que denominamos *Dipole Moment Content (DMC)*, una adaptación de la presentada por Morales et al.[3] basada en la noción de *dipole moment*, de la siguiente forma:

- t = tweets de un usuario
- pred(t), comunidad predicta por el modelo
- prob(t), probabilidad de la predicción
- $\upsilon(t) = \begin{cases} -prob(t) \text{ si pred(t)} = \text{C1} \\ prob(t) \quad \text{si pred(t)} = \text{C2} \end{cases}$
- $n^+ = (\upsilon(t) \ / \ \upsilon(t) > 0.9)$
  $n^- = (\upsilon(t) \ / \ \upsilon(t) < -0.9)$
- $\Delta A = \mid \frac{\#(n^+) - \#(n^-)}{\mid V \mid} \mid$
- gc^+ promedio de $\upsilon(t) \in n^+$
  gc^- promedio de $\upsilon(t) \in n^-$
- $\tau = \frac{\mid gc^+ - gc^- \mid}{2}$
- DMC = (1 - Δ A)τ

[3] https://developer.twitter.com/

[4] Q(G)=$\sum_{C \in G}(e_c - a_c)$, donde G es el grafo, C sus comunidades detectadas, e_c la fracción de ejes internos de la comunidad y a_c los de la frontera

[5] pasando su codificación a ASCII minúscula y removiendo los links, menciones, signos de puntuación y caracteres de control

[6] El entrenamiento lo realizamos con 5 epochs y un learning rate de 0.1[14]

El sentido de esta medida es que, cuanto más grande sea la diferencia de los promedios de predicción de cada comunidad ($\mid gc^{+} - gc^{-} \mid$), más alto será el score ya que mas diferenciable será la jerga de cada comunidad. Es importante notar que grandes diferencias en el tamaño de las comunidades (reflejadas en el valor de ΔA) tenderá a decrecer el valor total de la medida, lo que consideramos lógico dado que si una comunidad es mayoritaria ampliamente en al discusión el concepto de controversia se vuelve relativo.

3 Experimentos y Resultados

Realizamos nuestros experimentos en *datasets* de diversos contextos e idiomas, tomando los ya utilizados en [12] y agregando nuevos siguiendo el criterio definido por ellos para buscar temas polarizantes y no polarizantes. Por ejemplo para seleccionar temas controvertidos nos basamos en discusiones cubiertas por grandes medios de comunicación, que generaron un gran debate tanto on-line como off-line, como es el caso de la postulación de *Brett Kavanaugh* a la corte suprema estadounidense. Por otro lado para las discusiones no controvertidas seleccionamos temas relacionados a *soft news* o entretenimiento como por ejemplo el cumpleaños del artista chino *Jackson Wang*. Adicionalmente a estos criterios de selección de temas controvertidos o no, agregamos otro mecanismo de chequeo que es la visualización de la red a través del layout de Fruchterman & Reingold [11] en el cual pueden observarse a las comunidades con poca interacción entre sí distanciadas y más juntas en caso contrario. Esto lo hicimos para robustecer el criterio de búsqueda de discusiones controvertidas o no controvertidas.

En el siguiente beanplot se pueden ver los score obtenidos. El gráfico está divido en dos grupos, uno para discusiones con controversia y otro para aquellas sin controversia. Una gran separación de las dos distribuciones indica que la métrica tiene una buen rendimiento a la hora de caracterizar cada grupo. En este caso por ejemplo, dado que las medias están

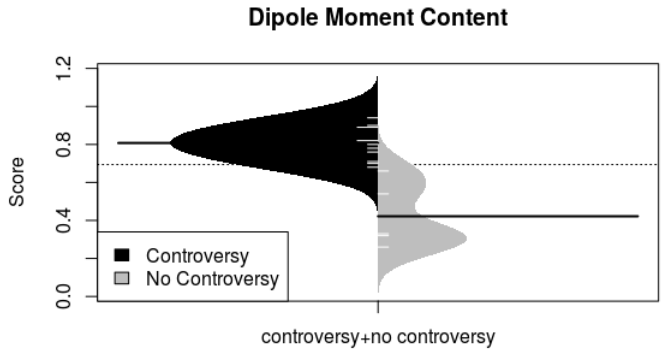

Fig. 1. Dipole Moment Content Scores sobre diversos grafos con y sin controversia

distanciadas y las varianzas no son muy grandes, podría establecerse un umbral de 0.67 a partir del cual puede indicarse que la discusión está polarizada. Establecemos este límite dado que el score mínimo obtenido dentro de los datasets con controversia es de 0.68 y el

máximo sin controversia 0.66. Mediante un análisis mas exhaustivo podríamos definir con mayor precisión y seguridad este valor, estos análisis preliminares lo que si nos confirman es que cuanto más cerca estén las métricas de la media de uno u otro conjunto (o incluso por debajo de la media de los sin controversia o por encima de la media de los controvertidos) mayor será la probabilidad de que la discusión pertenezca a dicho grupo.

En el siguiente beanplot se observan las métricas obtenidas por los distintos métodos desarrollados en [12]. Podemos ver que son dos los que logran una considerable diferenciación: EC y RWC. Ambos están basados únicamente en la estructura del grafo de retweets. Si bien las medias de cada grupo en estas métricas están mas separadas sus varianzas son mas grandes por lo que también poseen una zona de solapamiento entre las distribuciones de mediciones de ambos conjuntos, lo que resulta en una performance equiparable a nuestro método.

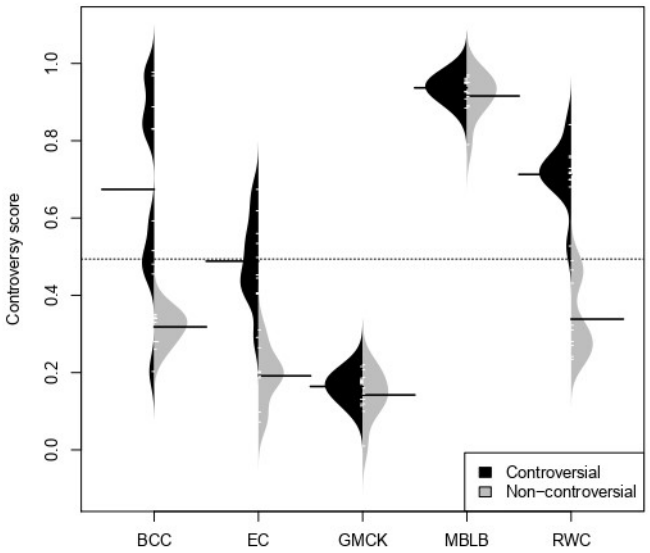

Fig. 2. Métodos desarrollados por Garimella et al. para detectar controversia

También es interesante notar que el rendimiento de la métrica *MBLB* es considerablemente inferior a la desarrollada por nosotros. La metodología de ambas es muy similar, la diferencia es que *MBLB* le asigna 1 y -1 al 5% de nodos mas apuntados de cada comunidad (autoridades) mientras que *Dipole Moment Content* lo hace a aquellos nodos etiquetados por nuestro modelo con una probabilidad mayor a 0.9.

4 Conclusiones y trabajo futuro

Mediante nuestro enfoque lingüístico logramos resultados equiparables a los obtenidos en [12] a través de enfoques estructurales, lo que nos indica que, a priori, es posible detectar la controversia en la jerga de las comunidades. A su vez, dada la mejora de nuestro método respecto de *MBLB* podemos señalar que a la hora de hablar de controversia el concepto de autoridad puede estar mejor definido a través del lenguaje que desde la estructura del grafo.

Adicionalmente ampliamos los datasets de prueba usados en [12] logrando así abarcar una gran diversidad de contextos y lenguajes como: Español, Inglés y Portugés, lo que aporta una considerable generalidad a nuestros resultados.

Si bien este es un trabajo aún preliminar, los resultados obtenidos son alentadores porque nos dan el indicio de que la controversia también se manifiesta en el lenguaje. Esto nos alienta a seguir mejorando y profundizando nuestros experimentos teniendo varios objetivos en el horizonte. En primer lugar ampliar más nuestros datasets de prueba para hacer más robusta y genérica la métrica y de esta forma confirmar si la polarización o no de una discusión se manifiesta de forma lingüística además de estructural. En segundo pensar otro método que nos permita hacerlo sobre datasets más chicos, ya que para poder entrenar nuestro modelo hace falta un cantidad mínima de tweets de alrededor de 10000. Para esto buscaremos utilizar otras técnicas como LDA [5] con la intención de detectar dos tópicos distintos y así poder diferenciar a los usuarios en las etapas 3 y 4. También probaremos otras técnicas de partición del grafo, como puede ser METIS [13], el método originalmente utilizado en [12].

Por último, a través de técnicas de interpretabilidad[8] podrían analizarse los modelos generados para ver como se caracterizan las jergas de cada uno de los lados de la discusión o que vocabulario es el que más separa una de otra.

Bibliografía

1. A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification
2. Abraham Doris-Down, Husayn Versee, and Eric Gilbert. 2013. Political blend: an application designed to bring people together based on political differences. In C&T. 120–130
3. AJ Morales, J Borondo, JC Losada, and RM Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos 25, 3 (2015).
4. Alexis Conneau, Holger Schwenk, Lo¨ıc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. arXiv preprint arXiv:1606.01781
5. BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. Journal of machine Learning research, 2003, vol. 3, no Jan, p. 993-1022
6. CALVO, Ernesto. Anatomía política de Twitter en Argentina. Tuiteando Nisman. Buenos Aires: Capital Intelectual, 2015.
7. DEL VICARIO, Michela, et al. Mapping social dynamics on Facebook: The Brexit debate. Social Networks, 2017, vol. 50, p. 6-16.
8. DOSHI-VELEZ, Finale; KIM, Been. A roadmap for a rigorous science of interpretability. arXiv preprint arXiv:1702.08608, 2017, vol. 150.
9. EASLEY, David, et al. Networks, crowds, and markets. Cambridge: Cambridge university press, 2010.
10. Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. 2013. Data portraits: Connecting people of opposing views. arXiv preprint arXiv:1311.4658 (2013).
11. FRUCHTERMAN, Thomas MJ, REINGOLD, Edward M. Graph drawing by force directed placement. Software: Practice and experience, 1991, vol. 21, no 11, p. 1129-1164.
12. da uno de los lados de la dis GARIMELLA, Kiran, et al. Quantifying controversy on social media. ACM Transactions on Social Computing, 2018, vol. 1, no 1, p. 3
13. George Karypis and Vipin Kumar. 1995. METIS - Unstructured Graph Partitioning and Sparse Matrix Ordering System. (1995).
14. HAYKIN, Simon; NETWORK, Neural. A comprehensive foundation. Neural networks, 2004, vol. 2, no 2004, p. 41.
15. Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P Gummadi, and Saptarshi Ghosh. 2015. Characterizing Information Diets of Social Media Users. In ICWSM.
16. KUMAR, Srijan, et al. Community interaction and conflict on the web. En Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018. p. 933-943.
17. Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In LinkKDD. 36–43.
18. Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In ICWSM.
19. Michael LaCour. 2012. A balanced news diet, not selective exposure: Evidence from a direct measure of media exposure. SSRN (2012)
20. PONS, Pascal; LATAPY, Matthieu. Computing communities in large networks using random walks. En ISCIS. 2005. p. 284-293.
21. STEWART, Leo G.; ARIF, Ahmer; STARBIRD, Kate. Examining trolls and polarization with a retweet network. En Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web. 2018.
22. Sean A Munson, Stephanie Y Lee, and Paul Resnick. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget.. In ICWSM.
23. Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. arXiv preprint arXiv:1502.01710.
24. Yoon Kim. 2014. Convolutional neural networks for sentence classification. In EMNLP.
25. ZACHARY, Wayne W. An information flow model for conflict and fission in small groups. Journal of anthropological research, 1977, vol. 33, no 4, p. 452-473.