ISSN 1870-4069

Análisis e implicaciones de la detección del humor en textos

Victor Manuel Palma-Preciado, Grigori Sidorov, Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación, México

{victorpapre, gr1965}@gmail.com,
{gelbukh}@gelbukh.com

Resumen. El presente trabajo tiene como objetivo analizar el fenómeno de la detección de humor en textos y aquellas implicaciones que conlleva, como puede ser el preparar los datos para tener un análisis correcto del set tomado, así mismo algunas metodologías actuales, todo esto tomando como base una selección de trabajos del periodo 2018-2019 los cuales hablan de la detección de humor, creación de conjunto de dato y talleres de detección de humor. Los trabajos fueron seleccionados dado el interés y los resultados que obtuvieron los autores, ya que entre ellos hay algunos con un enfoque más vanguardista. Como resultado, se obtuvo un análisis de los trabajos elegidos. Se puede observar que los métodos para detectar diferentes tipos de humor, aun teniendo la misma tarea por resolver, pueden ser variados y arrojar resultados competitivos uno frente a otros.

Palabras clave: Humorismo, detección de humor, conjuntos de datos, bromas.

Analysis and Implications of Humor Detection in Texts

Abstract. The paper aims to analyze a part of the phenomena of humor detection in texts and its implications, such as preparing the data to have a correct analysis of the set taken, as well as current methodologies. The analysis is based on a selection of works from the period 2016-2019, which talk about humor detection, creation of data sets and humor detection workshops. The works were selected given the interest and the results obtained by the authors, since among them there are some with a more avant-garde approach. As a result, an analysis of the chosen works was obtained. It can be seen that the methods for detecting different types of humor, even having the same task to be solved, can be varied and yield competitive results against each other.

Keywords: Humorism, humor identification, datasets, jokes.

1. Introducción

Muchas veces es difícil como humano identificar el humor, ya que a menudo el humor es subjetivo, dado que nos permite expresar un rango amplio de emociones,

normalmente es difícil para las personas identificarlo, ahora extrapolando la misma tarea, pero aun nivel computacional, se puede llegar a decir que es difícil identificar diferentes tipos humor y esto tiene cierto sentido, ya que el contexto nos dirá si por ejemplo, es una frase irónica o sarcástica, también de la misma forma si una frase intenta demostrar doble sentido o solo es una frase normal sin esa intención.

Todos esto problemas están presentes a la hora de identificar las intenciones tras de una frase y más aún cuando se tienen variantes de la lengua tomando por ejemplo los países de habla hispana donde ciertas palabras pueden tener un significado diferente o de connotación sexual que podría ser mal interpretado. Por lo tanto, se busca un modelo que nos permita identificar de manera positiva, si algo es gracioso o no y también que tan gracioso puede llegar a ser. Dado que siempre es bueno tener una noción general del humor y sus características, esta vez se analizarán textos concernientes a la creación de corpus y la detección del humor, por lo tanto, de este trabajo se puede esperar un análisis breve de ello.

La metodología usada se centró en compilar información relevante para la tarea de detección de humor en texto, en general se tomaron artículos y secciones de revistas así mismo algunos trabajos de Workshops donde el objetivo fuera la detección de humor. Seleccionado para el periodo 2016-2019, aunque solo se utilizó un trabajo del 2016 se deicidio considerar de estar forma el periodo de esta forma.

2. El concepto del humor

El humor tiene un rol importante en la sociedad, en este sentido [2], hablan de que el humor es una actividad comunicativa altamente inteligente, que puede provocar risas o asombro, pero sin la menor duda se puede decir que es un fenómeno sociológico. Si las máquinas llegarán a entender el idioma hasta cierto punto, les sería más fácil predecir la intención humana, lo que traería consigo una mejora en la capacidad que existe en la interacción humano-máquina. Sin embargo, [12], plantean la existencia de un problema en la detección del humor para el procesamiento del lenguaje natural [12]. La detección del humor a pesar de ser un fenómeno completamente conocido en la lingüística, aún presenta cierta carencia en las tareas computacionales [5,10].

Por lo tanto, el entender el humor se vuelve una tarea complicada al tener diferentes usos en la lengua, entre ellos los de tipo figurativo, metáforas, ironía y sarcasmo que son aspectos de la misma comunidad en la que vivimos y, por lo tanto, forman parte del acervo cultural escrito [3] Estos conceptos se utilizan de diferentes formas ingeniosas para expresar distintas ideas y esto pasa con mayor frecuencia en las redes sociales, en donde las frases o sus construcciones son más difíciles de entender, la creatividad y la forma en que se construye el humor permite visualizar las características personales de quien lo construye, por lo tanto, es importante crear nuevos métodos y recursos para analizar el humor en los textos [9].

3. Líneas base

Las líneas base (*baselines*) se definen como la base de comparación y se utilizan para evaluar el desempeño a través del tiempo. En el lenguaje natural cada autor toma

consideraciones diferentes para sus líneas base, algunas de las líneas base utilizadas como el caso de [1], tomaron dos consideraciones: en la primera, se consideró que la línea base fuera bajo una posibilidad del 50% de que el *tweet* se clasificará como humor o no y la segunda tuvo en cuenta la revisión del corpus, en el que la mayoría de los tweets humorísticos iniciaban con un guion, lo cual le daba precisión, pero con una exhaustividad (recall) menor, con una dificulta al identificar entre diferentes tipos de humor, como limitante de su modelo.

En el experimento de [9], se utilizó una escala para cada tweet del uno al cinco, en el cual, uno era el no gracioso y cinco como el más gracioso entre ese rango de números. Para medir los resultados se utilizó *Root Mean Squared Error* (RMSE). La *baseline* tomada fue justo en la mitad con tres puntos sobre el texto y su *baseline* sobre los datos de 1.14 RMSE. Todo esto bajo un corpus de 12,000 *tweets* en español, aunque se tuvieron diferentes variaciones del español esto no fue un impedimento para un post procesamiento de tweets consideradas de baja calidad, con las bases anteriores las tareas desarrolladas en el taller de *HAHA task* [1].

Mientras que las baseline usados por [11], utilizaron *Random Forest* con Word2Vec (W2V) adicionado de *Human Centric Feature* (HCF), optando por un *drop rate* de 0.5 y para testear el desempeño con factores F se considera el tamaño del filtro y HN.

Por otro lado, [6] decidieron utilizar teorías ampliamente reconocidas para el modelado del humor. Las líneas base que establecieron, se basaron en el trabajo hecho por [11], presentaron las características de sus modelos de la siguiente forma: estructura de incongruencia, la ambigüedad, el efecto de la interpersonalidad en la que las emociones juegan un papel, en la subjetividad que aumenta la posibilidad de que se genere el humor, el estilo fonético y como parte complementaria se utilizó *KNN Features*, construida bajo el indicador de similitud semántica.

Pero de la misma forma que [6] y también [2] utilizaron las baselines de [11], con otras ponderaciones para CNN, *Human Theroy driven Features* (HTF), *Human Centric Features* (HCF), para lo cual, los autores re implementaron el método propuesto por [11] más el uso de *KNN features*, Word2Vec similar a KNN características y por último HCF más Word2Vec.

4. Construcción de corpus

Se puede definir como un corpus a cualquier colección de piezas lingüísticas ordenadas de acuerdo a un criterio lingüístico para servir de muestra del lenguaje que se quiere estudiar. Se observa que los autores pueden desarrollar sus corpus [12] y por otro lado, existen aquellos que utilizan corpus ya marcados [1, 2, 9]

Al respecto, [12] crearon un conjunto de datos sobre el humor compuesto de 9123 chistes compilados de forma manual en el idioma chino. Sus anotaciones no solo contenían bromas, también incluían anotaciones hechas sobre el humor lingüístico, no solo aquellos que tuvieran una cierta cantidad de humor, también las palabras que desencadenaban el humor, sus relaciones, características y categorías humorísticas. Estos autores aseveran que el rápido crecimiento de las redes sociales, trae un número significativo de texto y dentro de estos los humorísticos. En la mayoría el humor se genera con dos conceptos incongruentes entre sí y examinándolo en un marco semántico, para saber si efectivamente ambos están semánticamente desconectados y

justamente validar la declaración anterior, se forma un contraste, que puede generar ambigüedad.

Por otra parte, [12] hablan de la importancia de los corpus, ya que en palabras sucintas un corpus es indispensable para el análisis de humor y determinara también la calidad de su detección automática. Por lo tanto, en lo que se basa la construcción de este grupo de datos, es en la premisa de los sets utilizados, para que expresen de forma adecuada cómo surge el humor. En general los datos recopilados fueron de twitter, en un esquema de anotaciones con palabras clave, que justamente hacían al texto gracioso, por la forma, la relación de los caracteres, el escenario, la categoría humorística y su grado humorístico.

[12] indican que trabajos previos no tenían en cuenta ciertas características en la detección del humor y su calidad. Las cuales hacen que su trabajo obtenga un esquema más completo, una de las principales diferencias obtenidas por estos autores fue la de anotar no solo lo que era humorístico sino también lo que causaba el humor.

El modelo desarrollado por los autores cuenta con las siguientes características de anotación:

JokeModel = (Relationship, Scene, Category, HumorLevel, Keyword, DataSource).

Se explica que cada tópico dentro del modelo maneja una conexión general, en el caso de la relación entre los actores puede ser el caso de doctor-paciente, amantes, subordinados, ya que obtener la relación ayuda a esclarecer el contexto de la broma, lo mismo pasara con la escena en la que se desarrolla. En el caso de la categoría y el nivel humorístico no hay consenso sobre qué se debe de usar, por lo tanto, los autores decidieron desarrollar sus propias categorías de humor y una anotación categórica de cinco niveles en los grados del humor.

Los autores determinan que las palabras claves resultan en uno de los desafíos más grandes, ya que muchas consideraciones tienen que ser tomadas, el leer el texto entero y establecer un significado general del texto, así como cada palabra del texto, para establecer su significado en el contexto que se utiliza, a su vez de estas mismas palabras determinan si su significado es de incongruencia, conflicto, ambigüedad, sorpresa o que desarrollen una emoción fuerte que haga al texto ser gracioso en el contexto dado, también decidir si el contexto en el que fue usada la palabra puede ser entendido.

El proceso que se tomó en cuenta para las anotaciones de los chistes se dio por validación cruzada, en el que solo si los anotadores estaban de acuerdo a la observación se consideraba completa. En general la fuerte necesidad de grupos de datos trae consigo nuevas formas de crearlos, nuevos parámetros que usar, así mismo trae consigo nuevas formas de interpretar los datos, como dicen los autores, es intrigante saber que la palabra con más alta frecuencia en ciertos chistes está sumamente ligada con la escena en la que son puestos, determinando su locación.

Por otro lado, [2] utilizaron corpus ya marcados, establecieron que para evaluar de manera correcta ocuparían un conjunto de datos que consistiría en muestras humorísticas (positivas) y no humorísticas (negativas).

Por lo tanto, su conjuntode datos consistirá de cuatro partes: *Puns of the Day* por [11], 16000 *One Liners* [7], *Short Jokes dataset* y *PTT jokes*, los cuales cuenta con bromas variadas, con diferentes tamaños y lenguajes.

En cambio, [9] en el preprocesamiento de sus datos, iniciaron con la limpieza de los símbolos de numeral que aparecen con mucha frecuencia en tweets, así como de

emoticons y *URL*, además de palabras usadas comúnmente en el ámbito de las redes sociales como los RT (*retweet*) y FAV (favoritos), estos últimos reemplazados por comodines que expresan el significado general de las palabras.

Se puede observar cierta tendencia a utilizar corpus con datos recabados de redes sociales o corpus con una amplia trayectoria de uso, se sobrentiende que no siempre será las mismas marcaciones para el corpus en el caso de que el modelo difiriera, ya que en raras ocasiones ocurrirá que se presente un corpus que esté exactamente marcado para el modelo que se vaya a utilizar. Se deben tomar ciertas consideraciones para ajustar la información, de tal forma que sea útil el marcaje y si esto no fuera suficiente, existe la opción de unir diferentes corpus para obtener uno que represente mejor nuestros datos.

5. Metodología para la detección del humor

La metodología para la detección del humor varía según el autor y la actividad que quiera desarrollar, pero en un enfoque general es utiliza *Deep learning*, aunque algunos autores llegan a utilizar machine learning de la forma de *Tree-CRF*.

Al respecto, [1] separan la información y evitan el sesgo de la mala escritura en twitter, este sesgo debe ser manejado de diferentes formas, entre ellas, mediante el análisis de sentimientos en forma de *Word-Vector* en árboles de *Condition Random Fields (Tree-CRF)* como clasificador de los datos, a su vez utiliza MalParser [8].

En el análisis de humor basado en anotaciones hechas por el humano (IberEval Workshop, 2018), en el cual las tareas requeridas fueron las siguientes: clasificar si un texto es humorístico o no y también predecir qué tan gracioso es. A partir de lo anterior, [9] propuso el uso de *Atention-based Recurrent Neural Network* (ARNN), donde la capa de cuidado o de atención ayude a calcular cada termino encaminado a encontrar las clases de humor.

Es importante recalcar que dichas actividades recaen sobre la información obtenida de Twitter, el cual, se volvió una plataforma popular para la obtención de contenido espontáneo creado por los usuarios.

También se nombran otros métodos para el reconocimiento del humor basados en aprendizaje supervisado, dentro de Deep Neuronal Network se encuentra *Long Short Term Memory* (LSTM) y también su variante bidireccional (Bi-LSTM). [9] hablaron del uso de Redes Neuronales Recurrentes (RNN) para la obtención de características para datos secuenciales. Estos autores utilizaron un ARNN, pero con la variante bidireccional de LSTM para generar el contexto del vector, lo cual, será pasado directamente a otra red LSTM para detectar si es humor o no, los autores describieron que dicho método no fue explorado en el campo de las arquitecturas basadas en ARNN para el reconocimiento del humor en español hasta ese momento.

[6] utilizó un modelo que tenía en tres características principales: complejidad métrica, humor en el texto y la expresión humorística del texto.

Además, midieron de diferente forma la complejidad de las oraciones, las cuales fueron cuantificadas como características respecto al número de sustantivos en frases, números de verbos, frases proposicionales, conjunciones subordinadas y otras características. Propusieron dentro de la estructura estadística, sus reglas de producción y por último las relaciones de dependencia, lo cual, indicó las relaciones entre palabras.

De igual forma, intentaron aplicar una combinación de POS *tag* y relación dependencia, obteniendo resultados poco favorables en comparación a solo utilizar relación de dependencia.

Como parte de la metodología que elaboraron [2], quienes utilizaron *Convolutional neural network* (CNN) diseñadas para extraer las características locales en dimensiones grandes de datos, entre ellas imágenes o conversaciones, quienes se apoyaron en una estructura basada en la tarea de clasificación de texto [4], lo primero fue tokenizar las sentencias de entrada de la forma word-vector con dimensiones D a una matriz de dos dimensiones, se utilizaron vectores GloVe, los cuales, fueron entrenados en token 6B, 400k vocabulario de Wikipedia 2014 más Gigaword 5 en la capa de *embedding*. También se utilizaron diferentes tamaños de filtros con un rango de tres a 20, para cada filtro de tamaño, aplicaron de 100 a 200 filtros, explotando el uso de *max pooling* y aplanando la salida, lo que conllevó a un vector aplanado 1D de dimensiones N en la salida predicha, se tomó en cuenta que se mejora el rendimiento bajo ciertas consideraciones, decidieron no enfatizar tanto la profundidad de la red, ya que esto traía consigo un incremento en la dificultad, tanto como sea profunda la red.

6. Análisis de modelos para detectar el humor

Como fue analizado, los enfoques con *Deep Learning* tienen resultados interesantes que permiten entender que este enfoque puede ser el adecuado para años venideros, sin olvidar los enfoques clásicos con los cuales comparar línea base.

Se exploraron características lingüísticas sobre el estilo por parte de [2], por ejemplo, número de palabras, número de caracteres usados, el énfasis por repetición y también el número de entre comillas. En cambio, para las características de estructura y contenido se usaron diferentes vocabularios como lo son de estilo topológico, animal, obsceno, sexual, además se tomó en cuenta la ambigüedad lexica (*Synsets*)(ADDESE). A su vez, se usaron algunas de las características afectivas, los sentimientos y el uso de emoticonos, para saber sí representaban una emoción negativa o positiva (*Emoticon Sentiment*).

Los datos superaron la línea base propuestos de 50% humor o no humor, se obtuvo un F_1 en su primera corrida de 0.7851, en la tercera y última de 0.7702. En general se puede observar que bajo el experimento desarrollado para las tareas propuestas los autores obtuvieron el mejor rendimiento en precisión (accuracy) en la primera fase frente a las demás, así mismo lograron una precisión y puntaje de F_1 más altos entre los contendientes, no de la misma forma en la exhaustividad (recall), donde tuvieron valores bajos.

Por otro lado, [6] adhirieron las características sintácticas estructurales en las líneas base propuestas, obtuvieron una mejoría notable en el rendimiento tanto del puntaje en precisión (accuracy) con un 7.9%, para el de puntaje de F_1 en un 7.2%, se puede apreciar que las características sintácticas pueden obtener mejores resultados con algunas propiedades del humor. Lo cual, significa que su modelo ayudo a identificar el humor y a explicar cómo estas estructuran sintácticas se relacionaron con el fenómeno lingüístico del humor. Postularon que, al identificar humor, los textos humorísticos utilizan palabras simples, pero estructuras sintácticas más complejas. Y a su vez, los textos tienden a parecerse más a conversaciones, no solo eso también este tipo de textos

son más vívidos y específicos, tienden a explicar que el texto deja al lector imaginar la situación y que ciertas palabras son utilizadas para realzar la situación que plantea.

Por otro lado, en la experimentación de [2] quienes decidieron optar por las líneas base [11], utilizaron *Random Forest* con Word2Vec (W2V) adicionado de Human *Centric Feature* (HCF), optaron por una taza de caída del 0.5 y para testear el desempeño con factores F y HN. El valor de F significa el incremento del tamaño del filtro, mientras que HN indica las capas del *Highway* para entrenar la red profunda, en este caso se eligieron tres capas ya que provee de mayor estabilidad y precisión (accuracy)en los pasos de entrenamiento. Se obtuvo un incremento en la puntuación F_1 de 0.859 a 0.903 con el modelo planteado usando CNN.

7. Conclusiones y trabajos a futuro

De los modelos analizados se aprecia que, en cuanto a la generación de corpus, la mejor fue la de [12] ya que su planteamiento de creación abarca características novedosas de marcado. En cambio, en la tarea de detección de humor [2,6] lograron modelos que mejoran las características de precisión (accuracy) y el puntaje de F_1. Los procesos con *Deep Learning* obtienen los resultados más favorables en la detección del humor, sin dejar pasar los enfoques clásicos.

En trabajos futuros se pretende extender para obtener un estado del arte y no un análisis pequeño como el que se puede observar en este artículo.

Agradecimientos. Agradecemos a CONACYT, SNI, IPN (SIP, COFAA), apoyo de proyectos SIP 20200859 y 20200797 y Conacyt A1-S-47854.

Referencias

1. Castro, S., Chiruzzo, L., Rosa, A.: Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval (2018)
2. Chen, P., Soo, V.: Humor recognition using deep learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2 (2018)
3. Ghosh, A., Veale, T.: Fracking sarcasm using neural network (2016)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1746–1751 (2014)
5. Kreuz, R., Glucksberg, S.: How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology (1989)
6. Liu, L., Zhang, D., Song, W.: Modeling sentiment association in discourse for humor recognition. Association of Computational Linguistics (2018)
7. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 531–538 (2005)
8. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables. In: NAACL-HLT, pp. 786–794 (2010)
9. Ortega-Bueno, R., Muñiz-Cuza, C.E., Medina-Pagola, J.E., Rosso, P.: UO-UPV: Deep linguistic humor detection in Spanish social media (2018)

10. Utsumi, A.: Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony 1. Journal of Pragmatics (2000)
11. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Conference on Empirical Methods in Natural Language Processing, pp 2367–2376 (2015)
12. Zhang, D., Zhang, H., Liu, X., Lin, H., Xia, F.: Telling the whole story: a manually annotated Chinese dataset for the analysis of humor in jokes (2019)