Minería de Emociones y Análisis Visual Aplicado a la Red Social Twitter

Florencia Marrocchi, Carolina Rapetti, Ana Maguitman, Elsa Estevez

Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Bahía Blanca, Argentina
{flor.marro, rapetticarolina}@gmail.com, {agm, ece}@cs.uns.edu.ar

Resumen. Las emociones forman un aspecto muy importante en nuestras vidas. Lo que hagamos o digamos reflejan de algún modo ciertas emociones. Analizar las emociones de un ser humano es fundamental para comprender su comportamiento, y puede lograrse a través de datos como texto, voz, expresiones faciales, entre otras. El cálculo de emociones consiste en la tarea de analizar texto para obtener emociones implícitas en él. Con la llegada de las redes sociales, analizar el texto disponible en estos sitios es una tarea muy atractiva. Analizar estos datos a través de Internet significa que estamos abarcando opiniones recientes de personas en distintos lugares del mundo sobre cualquier tópico deseado. Este documento resume un trabajo de investigación y desarrollo relacionado con la minería de emociones basado en el modelo emocional de Plutchik y enfocado en un escenario de gobierno digital. La contribución del mismo es presentar una extensión del modelo de Plutchik y mostrar una aplicación de software que lo aplica.

Palabras clave: Minería de Emociones; Análisis de Sentimientos; Rueda de Emociones de Plutchik; Gobierno Digital; Participación Ciudadana

1 Introducción

Una de las redes sociales con más usuarios en el mundo hoy en dia es Twitter, un servicio de microblogging que permite a sus usuarios enviar y publicar mensajes breves, generalmente solo de texto. Twitter es una fuente de información al instante, cualquier evento adquiere relevancia en cuestión de minutos, incluso las noticias se difunden antes que en cualquier medio periodístico. Según el estudio "Digital 2019: Q2 Global Digital Statshot" [1], en abril de 2019 esta red social contaba con unos 330 millones de usuarios activos. Su fortaleza se basa en la difusión de información en tiempo real, y en que es mayoritariamente pública, lo que la convierte en una fuente importante de datos de gran variedad. A partir de estos grandes volúmenes de datos, muchas organizaciones realizan escuchas sociales y monitoreos que si son procesados y transformados de manera correcta, pueden convertirse en una fuente de información útil para las mismas. El análisis de emociones es una de las opciones elegidas al momento de obtener una visión pública sobre un tema, producto o servicio en

particular, a partir de personas reales que están hablando del mismo de manera activa en línea.
El objetivo de este trabajo es extender el modelo de Plutchik y aplicarlo en una aplicación concreta. Para ello, se agregaron palabras en castellano a las ya provistas por el modelo y se desarrolló una aplicación web que permite generar una consulta online a Twitter según distintos tópicos de especial interés. A partir de la consulta, se obtienen los tweets en tiempo real, se realiza un análisis y clasificación de las emociones de los usuarios, y se las visualiza gráficamente. Finalmente, se utiliza la herramienta desarrollada para crear un escenario específico sobre gobernanza digital en el cual podamos visualizar y analizar los resultados obtenidos para un tema de debate en auge. Asimismo se incorporan características adicionales para incentivar al usuario, como la sugerencia de temas populares en Twitter en ese momento y una muestra de los tweets con los que se realizó la consulta.

El resto de este documento se estructura de la siguiente forma. La Sección 2 explica conceptos relacionados a minería de sentimientos, mientras que la Sección 3 presenta el modelo de Plutchik. La Sección 4 explica el diseño de la herramienta, y la Sección 5 las herramientas utilizadas para el desarrollo. La Sección 6 discute el uso de la herramienta en el contexto de las elecciones. Por último, la Sección 7 resume conclusiones y el trabajo futuro.

2 Conceptos Preliminares – Análisis de Sentimientos y Emociones

Para poder comprender lo narrado en este documento es necesario, primero, analizar temas de interés al trabajo realizado. En esta sección se tratará el tema "Minería de Emociones" aplicado al modelo de Plutchik [2], el cual sostiene que hay ocho emociones humanas primarias: alegría, confianza, miedo, sorpresa, tristeza, repugnancia, enojo y anticipación.

El procesamiento de lenguaje natural se utiliza entre otras cosas para el análisis de emociones y el análisis de sentimientos. Aunque el propósito de estos análisis sea similar, se diferencian en la identificación de los estados, por un lado el análisis de sentimiento detecta la polaridad del usuario, mientras que el análisis de emociones identifica específicamente las emociones que expresa un usuario según un modelo de emociones seleccionado.

Para realizar el análisis de emociones a través de texto, es necesario realizar un procesamiento de la información del mismo. Las técnicas que se emplean para diseñar e implementar un clasificador emocional se pueden clasificar en dos enfoques más una combinación de ellos. El primer enfoque es el basado en un lexicón. Un léxico emocional es un depósito de conocimiento que contiene unidades textuales anotadas con etiquetas emocionales. Se basan en recursos léxicos como léxicos, bolsas de palabras u ontología. El segundo enfoque es el basado en Machine Learning. Este enfoque utiliza algoritmos de aprendizaje automático para entrenar el sistema y mapear una función para la clasificación futura de emociones. Se basa en las características lingüísticas que elegimos para entrenar la máquina [3]. Este enfoque se

adapta a los cambios de dominio ya que aprende rápidamente nuevas características a partir del corpus (entrada de texto). El tercer enfoque es el híbrido. Este realiza una combinación de los dos primeros enfoques para obtener resultados de mayor precisión.

Para poder analizar las emociones en un texto, se debe hacer un preprocesamiento del mismo. El procesamiento previo reduce la dimensionalidad al eliminar los datos no deseados (ruido). El preprocesamiento en el texto se puede dividir en tres submódulos: tokenización, eliminación de ruido y normalización del texto, como se explica debajo.

En la *Tokenización* el texto se divide en tokens que representan palabras, oraciones o párrafos. La tokenización se realiza de acuerdo al limitador que separa una palabra, oración o párrafo de sí mismo. La división que genera la tokenización elimina las interrelaciones y los símbolos que le dan sentido al texto. En Twitter, por ejemplo, se consideran algunas situaciones especiales, ya que los tweets suelen contener menciones (nombre de usuario precedido por el símbolo '@') o links URL, los cuales deben ser considerados como tokens que serán removidos. Los tweets suelen contener también hashtags, pero a diferencia de los casos anteriores, éstos representan información importante a ser considerada, por lo que se conservarán como tokens y, por último, se hace la eliminación de *Stopwords*.

3 Modelo de Plutchik

Al momento de investigar sobre análisis de emociones en Twitter y herramientas públicas que permitan hacerlo, hemos encontrado que hay pocos trabajos que utilicen el idioma español. Por este motivo, se vio como una gran oportunidad crear una aplicación pública, que haciendo uso de tweets únicamente en español realice análisis de emociones para cualquier tema de interés que se mencione en dicha red social.

Por otra parte, este idioma generó algunas limitaciones. Fue necesario descartar el uso de varias herramientas para el procesamiento de lenguaje que no proporcionaban la capacidad de realizar operaciones sobre este idioma. También fueron descartadas fuentes de información que podrían servir para el análisis de emociones o para utilizar con alguna técnica de machine learning que únicamente hallamos en idioma inglés.

3.1 Extensión al Modelo

En función de la metodología elegida para el análisis (basada en lexicón), se requiere que los datos etiquetados estén en el idioma elegido. La fuente utilizada como lexicón originalmente se encontraba en idioma inglés, con una traducción automática al español, donde muchas de las palabras no se habían traducido correctamente.

El lexicón original (NRC Affect Intensity Lexicon) [4] se trata de una tabla excel con unas 14182 palabras, traducidas en 105 idiomas, donde cada una tiene diez atributos correspondientes a las ocho emociones de Plutchik más dos de polaridad. A

partir del lexicón inicial, se realizó el trabajo de traducir correctamente al español el conjunto dado de palabras, quitando las traducciones en los demás idiomas y dejando únicamente la lista de palabras con sus correspondientes 10 emociones. El lexicón final es entonces una traducción al español del NRC Affect Intensity Lexicon que quedará disponible como herramienta para futuros trabajos. Adicionalmente, se extendió el modelo con la clasificación de emojis, como se explica en la Sección 4.2.

4 Diseño de la Herramienta

El desarrollo de la aplicación web y el proceso de minería de texto para el análisis de emociones se realizaron utilizando los frameworks Angular [5] y Flask [6] respectivamente, siguiendo las siguientes etapas.

4.1 Búsqueda y Obtención de Tweets

La fuente de información que se utiliza en el proyecto son los tweets obtenidos de la base de datos de Twitter. Utilizando la librería Tweepy [7] se obtuvieron los tweets filtrando por idioma, palabras claves, fechas límite y cantidad de tweets. Además se incorporó la restricción de no obtener retweets, ya que estos implicarían información repetida que ya fue analizada. Una limitación de la versión gratuita de la API de Twitter [8] es que solo permite obtener tweets de hasta una semana atrás y limita la cantidad de tweets que se pueden obtener.

4.2 Incorporación de Emojis al Lexicón

Los emojis son símbolos muy utilizados en mensajes electrónicos y sitios web. Se trata de una representación de una emoción utilizando caracteres. Estos emojis representan una emoción en el tweet y es por eso que se decidió incluirlos en el análisis. Para ello se buscó el significado de cada emoji [9] y se incorporó al lexicón respetando el formato original. Finalmente, el lexicón usado quedó compuesto por todas las palabras iniciales y la unión de todos los emojis con sus respectivas emociones.

4.3 Pre-procesamiento de Texto

Con la lista de tweets obtenidos, se procedió a eliminar información que no fuera relevante para simplificar el análisis de emociones a realizarse por palabra, realizando:

Tokenización utilizando la librería NLTK [10], teniendo en cuenta los casos especiales como las expresiones propias de Twitter (hashtags, menciones y links URL).

Procesos no incluidos. Es una práctica común en las redes sociales utilizar mayúsculas cuando se quiere dar ímpetu a una expresión, como si se tratara de "aumentar" el tono. Sin embargo, se decidió ignorar esto y solo tener en cuenta el significado de la palabra para asignar la emoción correspondiente.

Otra situación común en los textos extraídos de redes sociales es que no respeta las reglas ortográficas del lenguaje. Sin embargo, hacer esta corrección podría desvirtuar la información obtenida transformando palabras que no están mal escritas, sino que simplemente no forman parte de la lengua española, lo que daría como resultado una emoción que no se corresponde con la palabra original.

Por último, la cantidad de retweets que tiene un tweet podría brindar información de cuántas personas están de acuerdo con ese contenido. Pero a su vez, Twitter permite escribir al retweetear un tweet ("citar tweet") por lo que es muy usual que los usuarios citen un tweet para explicar por qué no están de acuerdo con el mismo. Por este motivo, se decidió no obtener información a partir de la cantidad de retweets.

4.4 Análisis de Emociones

Habiendo finalizado el pre-procesamiento de los tweets, se tienen las listas de tokens para cada tweet, las cuales son procesadas utilizando el modelo de Plutchik mencionado anteriormente. La respuesta del procesamiento del texto (análisis de emociones) da como resultado entonces las emociones promedio para el conjunto de tweets obtenidos, las emociones para cada tweet, y los tweets utilizados para ese análisis. Con esta información se realiza la representación gráfica, como se explica en la siguiente sección.

4.5 Análisis Visual

La principal información a representar en la aplicación web consiste en las ocho emociones de Plutchik con sus respectivos promedios de apariciones encontrados en la búsqueda, para esto se utiliza un gráfico de radar, conformado por ocho emociones para los ejes, cada una representada con el color que le corresponde según Plutchik.

Adicionalmente se creó un gráfico de barras horizontal para representar la polaridad de la información obtenida, calculada de la misma forma que las emociones.

4.6 Aplicación Web

La aplicación realizada consiste de una app web single-page que contiene un formulario de búsqueda en el cual es posible ingresar los *hashtags a buscar*, donde el usuario tiene la posibilidad de ingresar hasta cinco hashtags de búsqueda. Luego de ingresar el primero, si así lo desea, tiene la posibilidad de adjuntar otro. La siguiente opción es la posibilidad de elegir si la búsqueda con más de un hashtag será realizada con un conjunto de tweets que contengan todos estos (conjunción) o un conjunto de tweets que contengan al menos uno (disyunción). En tercer lugar, dispone de la

posibilidad de limitar la fecha límite de búsqueda. Por último, debido a límites establecidos por la herramienta utilizada, hay un máximo de 3000 tweets que se pueden obtener en una consulta. La búsqueda se puede personalizar entonces para obtener resultados de entre 1 y 3000 tweets.

Además del formulario de búsqueda, se pueden ver dos gráficos de representación de información. El primero es un gráfico de tipo radar el cual muestra el porcentaje total de emociones obtenidas a partir de la búsqueda y, el segundo es un gráfico de barras horizontal que muestra la polaridad de los tweets obtenidos.

Una vez realizada la búsqueda, junto a los resultados de la misma se muestran los tweets, con el color de la emoción predominante en el mismo, a partir de los cuales se realizó el análisis.

Como característica adicional, se puede visualizar una lista con los 5 trending topics del momento, con la finalidad de sugerir al usuario temas de búsqueda.

Finalmente, ubicada en el menú principal de la aplicación, se encuentra la sección de ayuda, donde se explica brevemente cómo realizar una búsqueda.

4.7 Ejemplo de procesamiento de un tweet

Para comprender el proceso del análisis de emociones tomamos un tweet de ejemplo y analizamos cómo fue el procesamiento hasta llegar al resultado final.
Sea un tweet obtenido durante una búsqueda : "#CIOSummit2019 | Líderes de TI y de Transformación Digital compartirán experiencias de buenas prácticas e implementación de acciones que permiten una interacción entre los ciudadanos y el Estado más eficiente. ¡Estamos construyendo el legado! https://t.co/TS6MtcPayg #Colombia40 https://t.co/JBCUR8xSNp"
Los pasos realizados sobre el texto son:

1. Tokenización. El resultado obtenido es: "['líderes', 'de', 'ti', 'y', 'de', 'transformación', 'digital', 'compartirán', 'experiencias', 'de', 'buenas', 'prácticas', 'e', 'implementación', 'de', 'acciones', 'que', 'permiten', 'una', 'interacción', 'entre', 'los', 'ciudadanos', 'y', 'el', 'estado', 'más', 'eficiente', 'estamos', 'construyendo', 'el', 'legado']"
2. Búsqueda de cada token en el lexicón y obtención de clasificación.
3. Se realiza una sumatoria de resultados encontrados para cada emoción, cada token puede representar cero o más emociones: {'Positive': 2.0, 'Negative': 0.0, 'Anger': 0.0, 'Anticipation': 1.0, 'Disgust': 0.0, 'Fear': 0.0, 'Joy': 0.0, 'Sadness': 0.0, 'Surprise': 0.0, 'Trust': 1.0}
4. El mismo proceso se repite para todos los tweets obtenidos en la búsqueda. Finalmente se realiza un promedio de todos los valores de emociones encontradas según la cantidad de tweets analizados. Esta información es luego representada en un gráfico de radar.

5 Uso de la Aplicación en Gobierno Digital

La minería de emociones aplicada a una red social proporciona una gran oportunidad para analizar opiniones de usuarios con respecto a diferentes temas. El auge de Twitter en la actualidad y los diferentes debates de opinión que se generan en torno al contexto político del país, proporcionan una gran oportunidad para estudiar la influencia de la opinión de los usuarios sobre la imagen de los candidatos, las propuestas y el desarrollo de las campañas durante el período previo a las elecciones.

Entonces, realizar un análisis de emociones sobre las opiniones y comentarios que los usuarios de redes sociales expresan diariamente en relación a los hechos políticos que acontecen previo a las elecciones presidenciales, podría ser de gran utilidad para evaluar, en principio, de qué manera éstos influyen o afectan en la intención de voto o en el resultado de encuestas electorales oficiales, y posteriormente, determinar si se condicen con los resultados de las elecciones.

Se generaron diferentes escenarios relacionados al contexto político actual en Argentina cambiando los parámetros de entrada al sistema. Se mantuvo siempre el mismo rango de fechas para comparar en el mismo período de tiempo. Este rango se estableció desde una semana atrás a la fecha actual, esto es desde 10/07/2019 al 17/07/2019. También se mantuvo en 1000 la cantidad de tweets obtenidos.

5.1 Búsqueda de un Hashtag

#Elecciones2019. Luego de generar la búsqueda obtenemos como resultado el gráfico en la Figura 1 donde la emoción predominante es confianza, seguido de una mezcla de tristeza y anticipación.

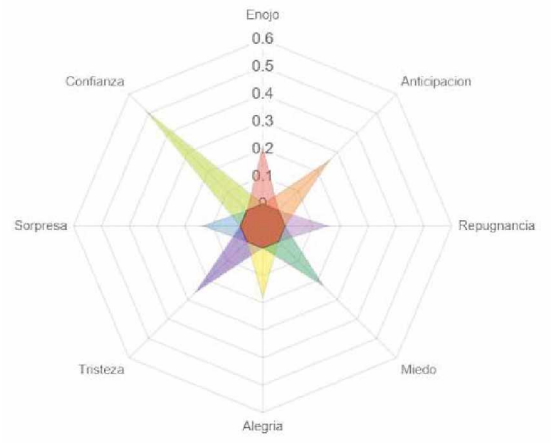

Fig. 1. Visualización de emociones con la búsqueda "#Elecciones2019".

5.2 Búsqueda de Hashtags en Disyunción (OR)

Mediante la función de agregar hashtags y la opción de disyunción, se realizaron búsquedas de dos o más hashtags donde los resultados obtenidos contengan alguno de los hashtags buscados. Los resultados dependen de qué tan relacionados están estos hashtags ya que son un promedio total de las emociones de los tweets obtenidos y si corresponden a temáticas no relacionadas es probable que se obtenga un resultado distorsionado. La búsqueda consistió de los hashtags "#cambiemos" o "#elecciones", donde se puede visualizar que la emoción predominante sigue siendo Confianza, pero crece el valor para Tristeza y disminuye para Anticipación de 0,3 a 0,2.

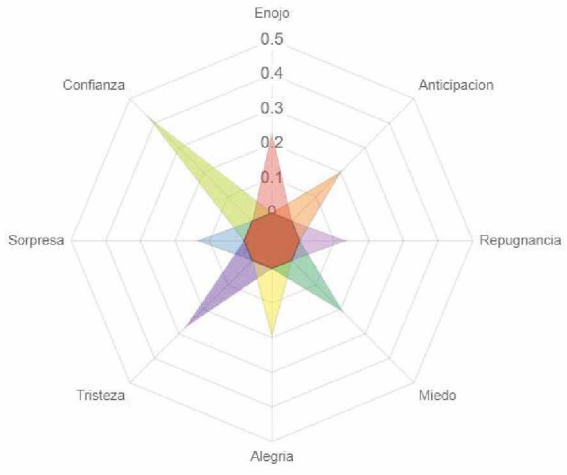

Fig. 2. Escenario búsqueda "#cambiemos" o "#elecciones".

En la Figura 2 se pueden visualizar ejemplos de diferentes tweets con su respectiva emoción predominante representada por el color del texto. En el primer texto la emoción que predomina es Enojo (color rojo del texto) y al leerlo se puede comprobar su concordancia. En el segundo caso el que el usuario que escribe utiliza la ironía, no contemplada en el análisis de emociones realizado, por lo cual la emoción predominante (confianza) no concuerda con la intención del escritor.

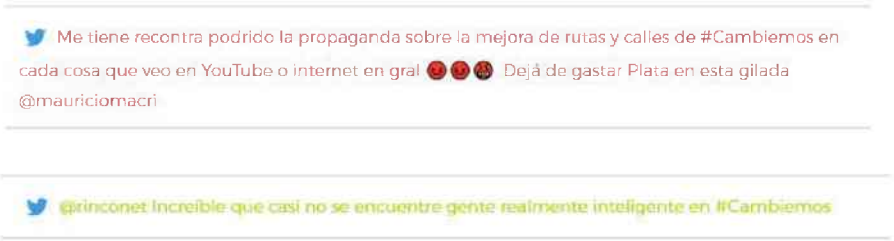

Fig. 3. Ejemplo de tweets con colores de emoción predominante.

5.3 Búsqueda de Hashtags en Conjunción (AND)

Se utilizó la función de agregar hashtags para realizar una búsqueda compuesta de hashtags con la opción conjunción. En la búsqueda realizada, "#cambiemos" y "#elecciones", se observa que la emoción predominante con gran diferencia sigue siendo Confianza, pero se suma la emoción Miedo, que se ubica apenas por debajo de Tristeza. Al analizar una conjunción, se puede afirmar que el resultado corresponde al conjunto, a diferencia de la disyunción, donde el resultado puede verse afectado por opiniones diferentes sobre cada uno de los temas individualmente.

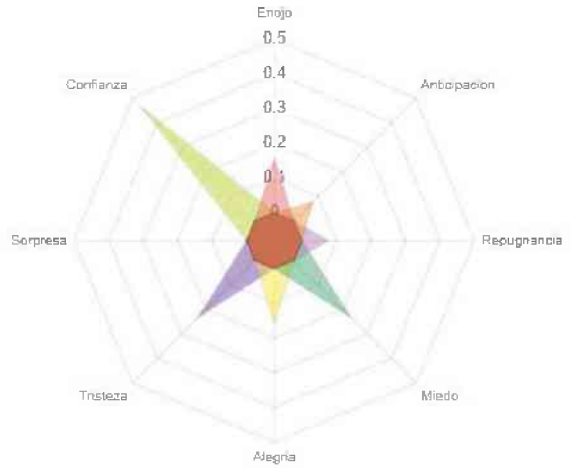

Fig. 4. Escenario de búsqueda "#cambiemos" y "#elecciones".

6 Conclusiones y Trabajo Futuro

Luego de haber finalizado las tareas propuestas, concluimos que se lograron los resultados esperados teniendo en cuenta que el análisis se hace sobre un lenguaje informal, con muchos errores de ortografía y abreviaciones que dificultan el procesamiento. Estos resultados se pueden mejorar implementando herramientas de lenguaje natural más complejas que profundicen el análisis. Por último, si bien se utilizó la aplicación considerando el escenario actual de las elecciones, la herramienta desarrollada permite analizar las emociones de los usuarios de Twitter sobre cualquier tema de interés para la sociedad.

Nuestro trabajo futuro se focaliza en las siguientes líneas. Una mejora e trabajar con la API paga de Twitter la cual permite acceso a tweets más antiguos y a mayor cantidad de los mismos. Así se podría realizar análisis en distintos períodos de tiempo y, por lo tanto, comparar un tópico en la actualidad con el mismo tópico en cierto período de tiempo atrás y analizar resultados. Otra mejora interesante a explorar es la

de incorporar técnicas de Machine Learning o aprendizaje supervisado. Durante el desarrollo del proyecto se buscó la forma de incorporar alguna técnica pero esto no fue posible ya que para aprendizaje automático, la técnica que más se adecuaba al proyecto, necesitaba un corpus de información etiquetada en español. Si bien existen corpus en idioma inglés, no fue posible encontrar un corpus en español para el enfoque de análisis de emociones adoptado en este trabajo. Esto representa una importante limitación a la hora de desarrollar métodos de aprendizaje supervisado en este idioma debido a que se trata de conjuntos de entrenamiento muy grandes que requieren ser traducidos "a mano" para no perder la esencia de las expresiones que pueden variar de un idioma a otro.

Referencias

[1] Kemp, S.: Digital 2019: Q2 Global Digital Statshot. DataReportal (2019). https://data reportal.com/reports/digital-2019-q2-global-digital-statshot [Accessed 17 Jul. 2019].
[2] R. Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci., vol. 89, no. 4, pp. 344–350 (2001).
[3] Moine, J., Haedo, A., Gordillo, S.: Estudio comparativo de metodologías para minería de datos. WICC (2011). https://digital.cic.gba.gob.ar/bitstream/handle/11746/ 3525/11746_3525.pdf-PDFA.pdf?sequence=1&isAllowed=y .
[4] Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 26–34 (2010).
[5] Angular, https://angular.io/.
[6] Flask, https://palletsprojects.com/p/flask/.
[7] Tweepy, https://www.tweepy.org/ .
[8] Twitter API, https://developer.twitter.com/en/products/tweets.
[9] Wood, I., Ruder, S.: Emoji as emotion tags for tweets. (2016). https://www.insight-centre.org/sites/default/files/publications/ianwood-emotionworkshop.p df .
[10] NLTK documentation, https://www.nltk.org/.