**PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ**

**ESCUELA DE POSGRADO**

**MAESTRÍA EN INFORMÁTICA**



**Análisis de sentimientos en reseñas de películas mediante el uso de Recursive Neural Tensor Networks.**

Curso : Técnicas avanzadas de minería de datos

Profesor : Cesar Beltrán

Integrantes : Gregory Cesar Valderrama Vilca

Diciembre 2015

Análisis de sentimientos en reseñas de películas mediante el uso de Recursive Neural Tensor Networks.

*Gregory Cesar Valderrama Vilca*

*Términos clave: Procesamiento de lenguaje natural, Aprendizaje profundo, Análisis de sentimientos, Redes neuronales, Aprendizaje profundo.*

# Introducción

En la actualidad existe una gran cantidad de documentos en la web que por su cantidad necesitan técnicas computacionales para clasificarles u organizarles.

Muchos de estos documentos no solo presentan información, sino que representan la opinión o interacción de los usuarios en distintos tópicos, es en este contexto que también necesitamos establecer la carga sentimental de un documento con el fin de saber si este es en alguna medida positivo o negativo. Este conocimiento puede ser aprovechado en distintas áreas del conocimiento como es la de Inteligencia de Negocios.

El análisis de sentimientos tiene un grado de complejidad mayor que la detección de tópicos en documentos pues requiere entender la semántica presente en las frases. Por mencionar un ejemplo esta expresión “*How could anyone sit through this movie?”* es claramente negativa, pero el análisis de términos independientes no proporcionaría una solución a este problema.

Debido a la complejidad para encontrar una respuesta en las técnicas tradicionales de aprendizaje del computador, es que las investigaciones más prometedoras en este campo han tomado la dirección del uso de redes neuronales en el contexto del llamado Aprendizaje profundo, debido a su capacidad para manejar modelos complejos para los cuales no se tiene una solución exacta.

En el presente trabajo utilizaremos las redes neuronales recursivas con el propósito de analizar la información de una página web que recopila opiniones sobre películas y les coloca un valor numérico del 1 al 5 dependiendo de la positividad en la reseña. El set de datos para nuestro experimento será obtenido de la página Kaggle de un concurso previamente realizado sobre el tópico de análisis de sentimientos.

En nuestro trabajo mostraremos el uso del kit de herramientas para el procesamiento del lenguaje natural en el contexto del análisis de sentimientos pudiendo obtener de un 58% a 64% de exactitud en dicho concurso.

En la primera parte de nuestro trabajo presentaremos el marco teórico que nos lleva a la utilización de las redes recursivas como una alternativa de solución al problema, en la segunda parte presentaremos las características de la data y el experimento, en la tercera parte presentaremos las conclusiones del experimento y trabajos futuros.

# Marco teórico

El análisis de la carga sentimental en documentos es una actividad que las personas podemos realizar con relativa simplicidad pero que es un área de gran desafío en el procesamiento de lenguaje natural entre otras cosas porque requiere de entender la semántica de las expresiones escritas.

Ha sido visto también como una forma de clasificación, por esta razón está relacionado a trabajos previos en la detección de géneros documentos con base en la información estadística [Karlgren, 1994] [Kessler, 1997], trabajos relacionados con detectar la subjetividad u objetividad también tienen cierta relación a nuestro tema de interés [Riloff, 2003]. Otros trabajos han abordado el problema de detectar la semántica de las frases y palabras usando heurísticas de lenguaje o pre seleccionados sets de palabras clave [Hatzivassiloglou, 1997] [Turney, 2002] que a su vez nos permiten indicar una polaridad de las mismas. Los trabajos también relacionados a la adquisición de “Hyponyms“ [Hearst, 1992] y el uso de lógica difusa para categorizar los documentos [Huettner, 2000] presentan algunos objetivos comunes con la detección de la polaridad en textos.

Trabajos directamente relacionados con el análisis de sentimientos y la polaridad semántica pueden ser encontrados en [Turney, 2002] y también en [Pang,2002].

Es pues el análisis de sentimientos un problema complejo para la clasificación automática de documentos, aun cuando pueda ser visto como una tarea de clasificación en dos categorías Positivo y Negativo; clasificadores como Naive Bayes, Maximum Entropy y Super Vector Machines, no pueden obtener grados de exactitud tan altos como los presentados en la categorización por tópicos de interés, cuyo grado de éxito es de 78% y 90% respectivamente [Pang,2002].

Distintas técnicas han sido utilizadas para mejorar los resultados con clasificadores no supervisados como el uso de uni-gramas, bi-gramas, partes del texto (adjetivos) y la posición de las sentencias en los documentos, pero los resultados obtenidos seguían en la misma proporción de alrededor 80% [Pang,2002].

Por ejemplo, sentencias como *“This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up”* o *“I hate the Spice Girls... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I’d despise every minute of it. But... Okay, I’m really ashamed of it, but I enjoyed it. I mean, I admit it’s a really awful movie ...the ninth floor of hell...The plot is such a mess that it’s terrible. But I loved it.”* Representan un real desafío a las técnicas de computador, aunque un ser humano puede fácilmente detectar la polaridad de las mismas.

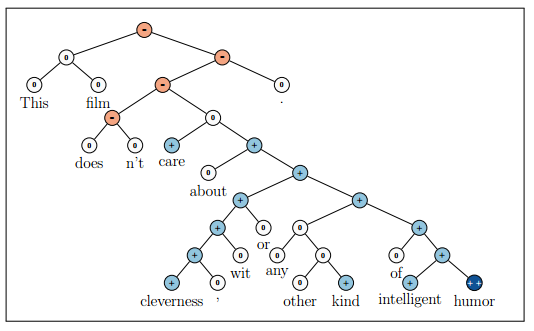
Según [Turney, 2010] los bajos resultados al momento de trabajar con esta complejidad, se dan por la imposibilidad de las estructuras computacionales de entender la semántica subyacente en el lenguaje humano. Por lo que hay necesidad de encontrar modelos computacionales que puedan representar no solo la existencia de términos sino la relación entre los mismos como son los modelos de espacios de vectores (Vector Space Models).

En la misma línea existen resultados en la detección de la semántica latente LSA aunque el hecho de utilizar las técnicas de matrices de documentos (Bag of words) presentan limitaciones al enfrentarse a casos como “house boat” y “boat house” porque en algunos contextos el significado es distinto, aunque el vector resultante será el mismo [Socher 2012].

Es así que la capacidad de analizar no solo los términos sino la forma en que se combinan los mismo es crucial para entender el lenguaje humano.

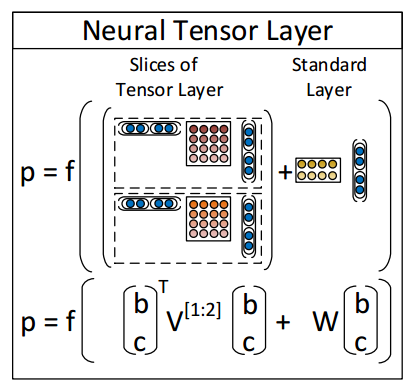
Por esta necesidad y basados en estudios previos con respecto al uso de las redes neuronales recursivas [Socher 2012], es que se proponen mejoras en el área de análisis de sentimientos que utilizan la semántica de las expresiones, no solo unarias, sino que también hacen uso de la información de su composición. Siendo el uso de la técnica Recursive Neural Tensor Network (RNTN) en el contexto más general de las técnicas de aprendizaje profundo, la que mejor precisión obtiene en comparación a trabajos anteriores con redes neuronales recursivas y técnicas de aprendizaje del computador como Naive Bayes o Super Vector Machines [Socher, 2013].

RNTN puede trabajar con vectores de cualquier tamaño, representando una frase mediante un vector y un árbol de expresiones, calculando los nodos superiores en base a los resultados de las hojas usando una función de composición. Como se muestra en la figura.



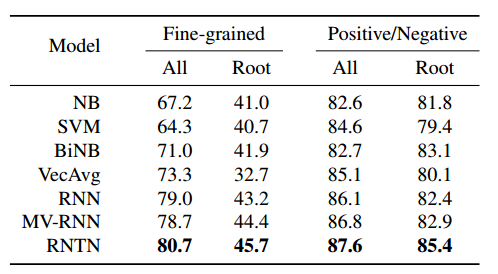
Además de obtener un alto grado de exactitud con respecto a las técnicas anteriores, presenta un buen desempeño al momento de enfrentarse a problemas de negación o cambio de orientación en la polaridad de la frase en los vocablos en ingles son representados por “but” y “not”.

Para solucionar el problema de un alto número de dimensiones que dependan del tamaño del vocabulario y mejorar la función de composición entre los nodos del árbol de términos es que [Socher, 2013] propone utilizar una misma función tensor para evaluar todos los nodos.



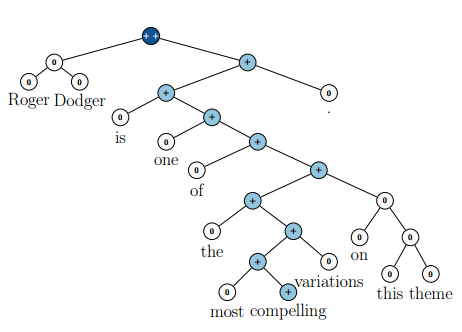
Para entrenar este tipo de redes utilizaremos un clasificador softmax tratando de maximizar la probabilidad de una correcta predicción o minimizar el grado de error según convenga.

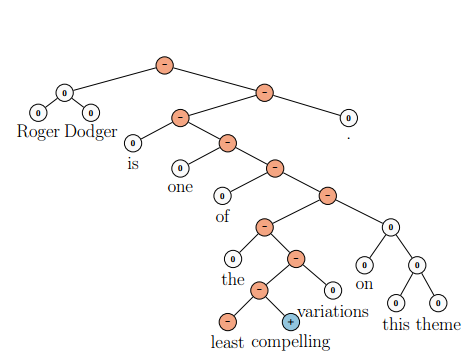
En los experimentos con RNTN se puede visualizar el siguiente resultado de exactitud comparado con diferentes técnicas. [Socher, 2013]

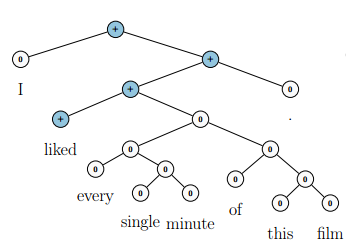


En un análisis más detallado se muestra que las redes RNTN presentan el mejor desempeño con respecto las expresiones de contraste (X but Y), reconocimiento de negación en expresiones positivas y en las expresiones de negación de la negación como “The movie was not terrible”.

A continuación, ejemplos de la evaluación recursiva en el árbol de términos para algunas frases. Siendo el color azul positivo el naranja el negativo y el blanco neutro.







# Experimento

El presente experimento trabaja sobre los datos obtenidos de un sitio web para la recolección de opiniones referentes a películas, http://www.rottentomatoes.com/, lo cual lo hace un dominio especialmente útil para el análisis de sentimientos, pues asociada a cada reseña existe una polarización indicada por el número de estrellas asociadas a la película.

Aunque la data es de fácil extracción este dominio de problema es bastante útil como desafío para las técnicas computacionales, debido a que puede contener términos con ambivalencias y sarcasmos.

El set de datos para nuestro experimento será obtenido del sitio web <https://www.kaggle.com/> en la competencia de análisis de sentimientos, que consta de datos para el entrenamiento y prueba del algoritmo de clasificación, así como en el sitio web se puede obtener un resultado sobre la exactitud del modelo entrenado y una escala de éxito con respecto a intentos anteriores.

Para acceder al código fuente y modelos entrenados se puede acceder al siguiente proyecto en GitHub el cual puede ser descargado libremente. <https://github.com/gcvalderrama/pucp_dl_RNTN>

Nuestro trabajo hará uso del kit de herramientas para el procesamiento de lenguaje natural de la universidad de Stanford [Manning, 2014].

Nuestro corpus está conformado de datos de entrenamiento que tienen un identificador de sentencia, la sentencia y el sentimiento asociado a dicha expresión en base a la siguiente escala.

* 0 negativo
* 1 algo negativo
* 2 neutral
* 3 algo positivo
* 4 positivo

El corpus de entrenamiento estará conformado de 8530 sentencias que en conjunto tendrán 156060 frases cada una anotada con el valor del sentimiento asociado. Con la siguiente distribución.

La data para el reto en Kaggler tiene 3310 sentencias con un total de 66291 frases las cuales deben ser también categorizadas en base a la escala previamente expuesta.

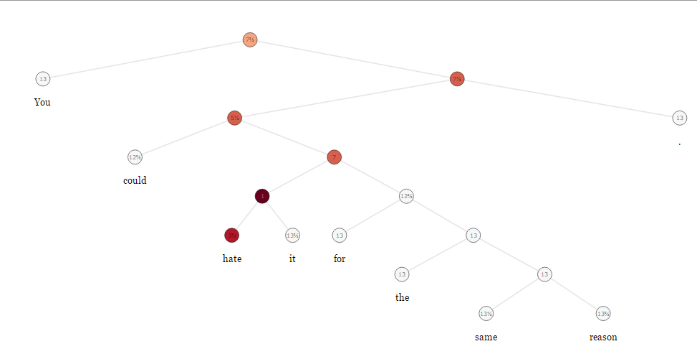
En nuestros experimentos utilizaremos la RNTN que fue propuesta en el trabajo de [Socher, 2013], por esta razón necesitamos transformar la data a un estilo de formato conocido como Treebank más apropiado para interactuar con el algoritmo propuesto que hace uso de la estructura de árbol presenten en las sentencias.

Se respetarán también la información de mayúsculas y caracteres especiales por tratarse del dominio de películas y sabiendo que estas pueden tener nombres con reglas especiales.

Por ejemplo, para la siguiente expresión debemos conseguir un resultado similar a:

“You could hate it for the same reason”

“(1 (1 You) (1 (0 (2 could) (1 (0 (0 hate) (1 it)) (2 (1 for) (2 (1 the) (2 (2 same) (2 reason)))))) (1 .)))”



Como podemos apreciar ahora a cada palabra de la frase le corresponde un valor de sentimiento, así como a la combinatoria entre las mismas también les corresponde un valor de sentimiento.

Es pues en la data para el aprendizaje que se requiere de mucha más exactitud y cuidado pues dependiendo de este conocimiento la red neuronal podrá aprender mejor las reglas de combinación sobre los diferentes términos.

Con esta información y separando un 5% de los datos como set de desarrollo y pruebas para medir el avance de nuestro aprendizaje, se procedió a iniciar los ciclos de aprendizaje.

Nuestro primer experimento fue utilizar la data de entrenamiento y modelo propuesto originalmente por la universidad de Stanford el cual alcanza un grado de exactitud mayor al 80% en su data de entrenamiento, esto es posible porque tanto la universidad como Kaggler utilizan el mismo dominio de problema, como son las reseñas de películas obtenidas del mismo sitio web [Socher, 2013].

Con este experimento obtuvimos 64.5% de exactitud en la página web de Kaggler, lo cual nos prueba la influencia de la data de entrenamiento en el modelo resultante. [Socher, 2013]. Como se menciona los investigadores de la universidad de Stanford tuvieron mucho cuidado en obtener una data de entrenamiento correctamente categorizada en cuanto a la carga de sentimientos se refiere. A continuación, mostramos el resultado de exactitud, en el Anexo del experimento número uno se encontrará la matriz de confusión y el grafico que resume la cantidad de éxito y distancia de error con respecto a nuestro set de validación.

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.372148 | 0.91 |
| Exactitud Positiva | 0.440451 | 0.95 |
| Promedio | 0.408178 | 0.93 |

De los resultados podemos encontrar que hay 7 casos con un error extremo, que corresponden por ejemplo a esta expresión. *“this is the opposite of a truly magical movie.”*, donde el sentimiento esperado es muy negativo pero el resultado fue muy positivo. Lo que valida la dificultad de entender la semántica del sarcasmo.

Nuestro siguiente experimento fue entrenar el modelo RNTN con la data de entrenamiento de la página web Kaggler, después de 400 épocas en al rededor de 6 horas se alcanzó un modelo de entrenamiento que para nuestra data de validación mostraba la siguiente tabla de exactitud.

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.460661 | 0.63 |
| Exactitud Positiva | 0.654686 | 0.7 |
| Promedio | 0.563011 | 0.665 |

Al evaluar el modelo en el reto de Kaggler obtuvimos un resultado de exactitud 58.165 %. Para un mayor detalle sobre la matriz de confusión para este experimento podemos utilizar el anexo 2.

En este experimento podemos ver que la exactitud y sensibilidad de los resultados ha decaído, en nuestro análisis hemos podido notar que la asignación de sentimientos a la data de entrenamiento parece ser menos exacta que la proporcionada por la universidad de Stanford lo cual puede estar afectando el entrenamiento.

Nuestro tercer experimento consistirá en incrementar la cantidad de datos de entrenamiento fusionando la data de entrenamiento proporcionada por la universidad de Stanford y la data proporcionada por Kaggler.

Después de entrenar por 373 épocas y aproximadamente 24 horas de entrenamiento el modelo muestra la siguiente exactitud.

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.357592 | 0.79 |
| Exactitud Positiva | 0.446794 | 0.83 |
| Promedio | 0.404647 | 0.81 |

Al evaluar el modelo en Kaggler se obtuvo un 61.63% de exactitud, lo cual significa que incrementamos nuestro resultado en 3% con respecto al resultado anterior, pero, aunque los casos de error extremo han desaparecido no alcanzamos la exactitud del resultado del primer experimento.

# Conclusiones y trabajos futuros

El análisis de sentimientos en textos es un tema de gran desafío en el área del procesamiento de lenguaje natural pues necesita de entender la semántica de las composiciones de palabras y como estas varían dependiendo de su posición o relación con otras en una misma sentencia.

De nuestros experimentos podemos concluir que el uso de técnicas de aprendizaje profundo es una solución a un problema con este grado de complejidad donde una solución exacta parece muy lejana en este momento.

La forma en que las RNTN pueden manejar las relaciones de combinación entre términos de una misma expresión representan muchas posibilidades en distintos ámbitos relacionados al procesamiento del lenguaje natural.

De los experimentos realizados también podemos concluir que la base de extracción de lemas, la normalización y la anotación de términos son tareas importantes pues de ellas obtendremos más información que podremos explotar en el análisis de la semántica de las sentencias.

El adecuado modelamiento del conocimiento que otorgaremos a las redes neuronales, por ejemplo, en nuestro caso la adecuada valoración sentimental de la combinatoria de los términos en una sentencia, representan un componente fundamental para obtener un buen modelo de aprendizaje, de nuestros experimentos también podemos ver que el incremento del volumen de datos de entrenamiento no impacta significativamente en la capacidad de aprendizaje, siendo mas relevante la idoneidad de los datos referentes a la valoración sentimental.

Es pues un trabajo pendiente el poder identificar técnicas que nos permitan identificar y extraer conocimiento valido para así poder entrenar de la manera mas adecuada los modelos de redes neuronales de aprendizaje profundo.

# Referencias

[Pang,2002] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

[Karlgren, 1994] Karlgren, Jussi, and Douglass Cutting. "Recognizing text genres with simple metrics using discriminant analysis." Proceedings of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1994.

[Kessler, 1997] Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze. "Automatic detection of text genre." *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997.

[Riloff, 2003] Riloff, Ellen, and Janyce Wiebe. "Learning extraction patterns for subjective expressions." *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003.

[Hatzivassiloglou, 1997] Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives." *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*. Association for Computational Linguistics, 1997.

[Turney, 2002] Turney, Peter, and Michael L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus." (2002).

[Hearst, 1992] Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992.

[Huettner, 2000] Huettner, Alison, and Pero Subasic. "Fuzzy typing for document management." *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes* (2000): 26-27.

[Turney, 2010] Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." *Journal of artificial intelligence research* 37.1 (2010): 141-188.

[Socher, 2012] Socher, Richard, et al. "Semantic compositionality through recursive matrix-vector spaces." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

[Socher, 2013] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vol. 1631. 2013.

[Manning, 2014] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. [The Stanford CoreNLP Natural Language Processing Toolkit](http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf). In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Anexos

## Anexo 1 - Primer experimento Datos de Entrenamiento Stanford

|  |  |  |
| --- | --- | --- |
| **-** | **Frases** | **Sentencias** |
| Total | 8898 | 250 |
| Correctas | 4843 | 141 |
| Incorrectas | 4055 | 109 |
| Exactitud | 0.54428 | 0.564 |

### Matriz de confusión para las frases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg. (Guess)** |
| 0 | 58 | 15 | 5 | 0 | 1 | 79 |
| 1 | 230 | 643 | 162 | 29 | 15 | 1079 |
| 2 | 662 | 838 | 3158 | 868 | 675 | 6201 |
| 3 | 29 | 54 | 187 | 840 | 221 | 1331 |
| 4 | 6 | 7 | 6 | 45 | 144 | 208 |
| Marg. (Gold) | 985 | 1557 | 3518 | 1782 | 1056 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 0.73418 | 0.05888, | 0.99735, | 0.10902 |
| 1 | 0.59592 | 0.41297, | 0.94061, | 0.48786 |
| 2 | 0.50927 | 0.89767, | 0.43439, | 0.64986 |
| 3 | 0.6311 | 0.47138, | 0.931, | 0.53967 |
| 4 | 0.69231 | 0.13636, | 0.99184, | 0.22785 |

### Matriz de confusión para las sentencias

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg. (Guess)** |
| 0 | 11 | 2 | 1 | 0 | 0 | 14 |
| 1 | 36 | 42 | 19 | 4 | 0 | 101 |
| 2 | 0 | 3 | 19 | 1 | 0 | 23 |
| 3 | 3 | 3 | 11 | 42 | 23 | 82 |
| 4 | 0 | 0 | 0 | 3 | 27 | 30 |
| Marg.(Gold) | 50 | 50 | 50 | 50 | 50 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 0.78571 | 0.22 | 0.985 | 0.34375 |
| 1 | 0.41584 | 0.84 | 0.705 | 0.55629 |
| 2 | 0.82609 | 0.38 | 0.98 | 0.52055 |
| 3 | 0.5122 | 0.84 | 0.8 | 0.63636 |
| 4 | 0.9 | 0.54 | 0.985 | 0.675 |

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.372148 | 0.91 |
| Exactitud Positiva | 0.440451 | 0.95 |
| Promedio | 0.408178 | 0.93 |

## Anexo 2 – Segundo Experimento Datos de entrenamiento Kaggler

|  |  |  |
| --- | --- | --- |
| **-** | **Frases** | **Sentencias** |
| Total | 8898 | 250 |
| Correctas | 3356 | 96 |
| Incorrectas | 5542 | 154 |
| Exactitud | 0.377 | 0.38 |

### Matriz de confusión para las frases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg.(Guess)** |
| 0 | 84 | 58 | 88 | 13 | 5 | 248 |
| 1 | 395 | 634 | 752 | 318 | 189 | 2288 |
| 2 | 104 | 330 | 1542 | 364 | 91 | 2431 |
| 3 | 390 | 508 | 1021 | 949 | 624 | 3492 |
| 4 | 12 | 27 | 115 | 138 | 147 | 439 |
| Marg.(Gold) | 985 | 1557 | 3518 | 1782 | 1056 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 0.33871 | 0.08528 | 0.97927 | 0.13625 |
| 1 | 0.2771 | 0.40719 | 0.77469 | 0.32978 |
| 2 | 0.63431 | 0.43832 | 0.83476 | 0.51841 |
| 3 | 0.27176 | 0.53255 | 0.64264 | 0.35988 |
| 4 | 0.33485 | 0.1392 | 0.96276 | 0.19666 |

### Matriz de confusión para las sentencias

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg.(Guess)** |
| 0 | 19 | 1 | 1 | 1 | 0 | 22 |
| 1 | 19 | 24 | 20 | 8 | 6 | 77 |
| 2 | 6 | 14 | 11 | 10 | 5 | 46 |
| 3 | 5 | 11 | 18 | 27 | 24 | 85 |
| 4 | 1 | 0 | 0 | 4 | 15 | 20 |
| Marg.(Gold) | 50 | 50 | 50 | 50 | 50 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 0.86364 | 0.38 | 0.985 | 0.52778 |
| 1 | 0.31169 | 0.48 | 0.735 | 0.37795 |
| 2 | 0.23913 | 0.22 | 0.825 | 0.22917 |
| 3 | 0.31765 | 0.54 | 0.71 | 0.4 |
| 4 | 0.75 | 0.3 | 0.975 | 0.42857 |

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.460661 | 0.63 |
| Exactitud Positiva | 0.654686 | 0.7 |
| Promedio | 0.563011 | 0.665 |

## Anexo 3 – Tercer experimento Datos de entrenamiento de Stanford y Kaggler

|  |  |  |
| --- | --- | --- |
| **-** | **Frases** | **Sentencias** |
| Total | 8898 | 250 |
| Correctas | 4887 | 118 |
| Incorrectas | 4011 | 132 |
| Exactitud | 0.5492 | 0.47 |

### Matriz de confusión para las frases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg.(Guess)** |
| 0 | 73 | 5 | 7 | 0 | 1 | 86 |
| 1 | 199 | 632 | 156 | 46 | 19 | 1052 |
| 2 | 677 | 856 | 3169 | 822 | 682 | 6206 |
| 3 | 32 | 57 | 177 | 861 | 202 | 1329 |
| 4 | 4 | 7 | 9 | 53 | 152 | 225 |
| Marg.(Gold) | 985 | 1557 | 3518 | 1782 | 1056 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 0.84884 | 0.07411 | 0.99836 | 0.13632 |
| 1 | 0.60076 | 0.40591 | 0.94279 | 0.48448 |
| 2 | 0.51063 | 0.9008 | 0.4355 | 0.65179 |
| 3 | 0.64786 | 0.48316 | 0.93423 | 0.55352 |
| 4 | 0.67556 | 0.14394 | 0.99069 | 0.23731 |

### Matriz de confusión para las sentencias

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Guess/Gold** | **0** | **1** | **2** | **3** | **4** | **Marg.(Guess)** |
| 0 | 5 | 0 | 0 | 0 | 0 | 5 |
| 1 | 39 | 35 | 17 | 5 | 2 | 98 |
| 2 | 6 | 9 | 20 | 6 | 4 | 45 |
| 3 | 0 | 6 | 12 | 37 | 23 | 78 |
| 4 | 0 | 0 | 1 | 2 | 21 | 24 |
| Marg.(Gold) | 50 | 50 | 50 | 50 | 50 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **-** | **Precision** | **Recall** | **Spec** | **F1** |
| 0 | 1 | 0.1 | 1 | 0.18182 |
| 1 | 0.35714 | 0.7 | 0.685 | 0.47297 |
| 2 | 0.44444 | 0.4 | 0.875 | 0.42105 |
| 3 | 0.47436 | 0.74 | 0.795 | 0.57812 |
| 4 | 0.875 | 0.42 | 0.985 | 0.56757 |

|  |  |  |
| --- | --- | --- |
| **-** | **Frase** | **Sentencia** |
| Exactitud Negativa | 0.357592 | 0.79 |
| Exactitud Positiva | 0.446794 | 0.83 |
| Promedio | 0.404647 | 0.81 |