Relation Classification: Análisis del efecto de la proporción de relaciones positivas y negativas

Álvaro Dueñas Fernandez

Universidad del País Vasco (December 14, 2020)

Objetivos

- Objetivo: observar los resultados obtenidos en función de las distintas proporciones entre relaciones negativas y positivas.
- Preguntas de la investigación:
  RQ1 ¿Cual es la proporción que mejores resultados arroja?
  1. 1+ : 1-
  2. 1+ : 2-
  3. 1+ : 3-
  4. 1+ : 4-

  RQ2 ¿Cual es el mejor clasificador?
  1. Baseline(RandomForest)
  2. SupportVectorMachine

Preproceso y Datos

Preproceso

- Lectura de los datos, tokenizar los datos y representarlos con un vector.
- Fuente de datos: [Bossy et al., 2019]
- Class = Tipo de Relación. Siendo nula para las relaciones negativas y las posibles positivas son LivesIn y Exhibits.

Representación: se usa una representación vectorial del texto y entidades envueltas.

Representación de textos

- **Pre-Proceso** Validar los datos, tokenizarlos, generar las relaciones negativas y vectorizarlos.
- **Representación** Para representar los textos se ha hecho uso del modulo tfidf de sklearn [Pedregosa et al., 2011] y doc2vec de gensim [Řehůřek and Sojka, 2010].

Prop	Num Rel	tfidf	doc2vec
1:1	2251	3720	3315
1:2	3373	3720	3477
1:3	4492	3720	3476
1:4	5598	3720	3706

Classifier 1: baseline

Se ha optado por RandomForest como baseline porque es capaz de manejar valores negativos(correspondientes a los embeddings). Consiste en un conjunto de arboles de decisión con los que la clase a predecir es decidida por mayoría.

Classifier 2: (Support Vector Machine)

Es un clasificador ideal para trabajar con textos, mediante combinaciones lineales o polinómicas de los atributos de las instancias segrega la clase predicha.

Resultados experimentales para validar RQ1

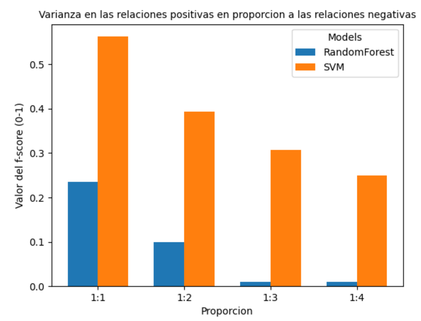

Figure 1:Avg Fscore de las clases + para cada proporción tfidf.

- Más compleja la clasificación de relaciones + a medida que es mayor la proporción. El modelo SVM es más adecuado para realizar esta tarea, ya que arroja mejores resultados.

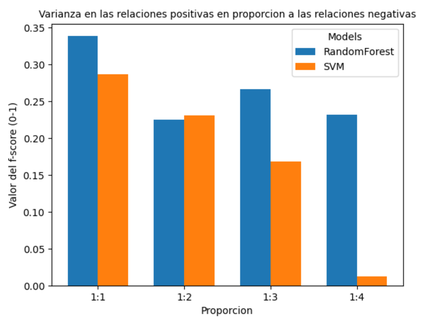

Figure 2:Avg Fscore de las clases + para cada propor doc2vec.

- Los embeddings no son una representación adecuada para clasificar con el modelo SVM.

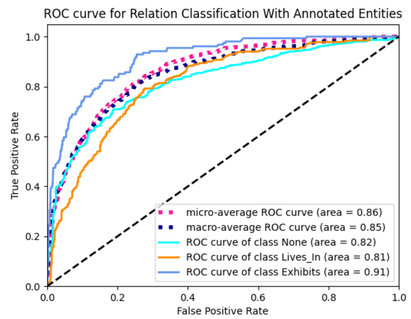

Figure 3:Curva Roc en condiciones optimas. tfidf

- El modelo consigue predecir de manera efectiva en el mejor de los casos. Y es mejor a la hora de clasificar las relaciones de clase Exhibits.

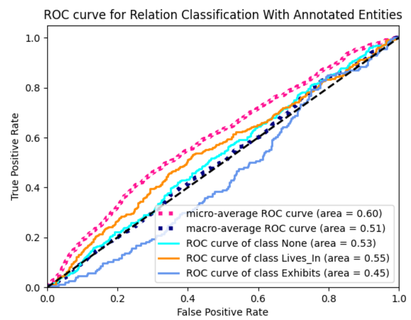

Figure 4:Curva Roc en condiciones optimas. doc2vec

- Los embeddings no son una representación adecuada para clasificar con el modelo SVM.

Experimental results: RQ2

A continuación se muestra el heatmap de la mejor combinación de parámetros posibles.

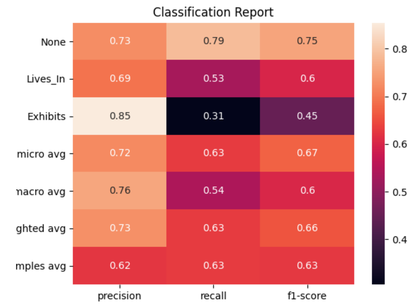

Figure 5:Extra results to validate your hypothesis

Corresponde a una representación tfidf, con el mínimo de proporción posible a partir del modelo SVM.

Discusión

- Los resultados muestran la complejidad de identificación de una minoría.
- Los resultados están dentro de los umbrales esperados.

Conclusiones

- Es necesaria una optimización exhaustiva para no decaer en la identificación de las clases positivas.
- Fortalezas: Es un software muy ágil y analiza con diferentes métricas los resultados.
- Debilidades: Carencia de variedad de clasificadores.

Bibliography

[Bossy et al., 2019] Bossy, R., Deléger, L., Chaix, E., Ba, M., and Nédellec, C. (2019).
Bacteria biotope at bionlp open shared tasks 2019.
pages 121–131.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).