# Informe de Resultados — Trabajo Práctico 1

**Materia: Aprendizaje Automático Avanzado**

Profesores: Dr. Juan Santos, Lic. Elimiano Churruca

Alumnos: Franco Serafini, Seivane Nicolás, Matías Cisnero

## Resumen general

### Métricas Utilizadas

Para la realización del siguiente informe se utilizará la métrica *Accuracy*, la cual representa el porcentaje de predicciones correctas respecto al total de predicciones realizadas. En este caso, se introduce una ligera modificación en su aplicación. Para la predicción de la palabra central, se considera el valor con mayor probabilidad obtenido por la función *soft-max* al momento de la inferencia, lo que corresponde a la métrica **Top1**. Asimismo, se evaluará la métrica **Top5**, que toma en cuenta los cinco valores más altos devueltos por la función *soft-max*, considerando una predicción correcta si la palabra esperada se encuentra dentro de esos cinco primeros resultados.

$$\text{Accuracy} = \frac{\text{Número de predicciones correctas}}{\text{Número total de muestras}} \tag{1}$$

### Corpus sin Fragmentar

**Características del corpus:** Se realizó un método de obtención de *Representaciones contextuales, CBOW*, donde se tomo a cada palabra y signo de puntuación como entrada del vocabulario, dando un total de 310.347 palabras en el corpus y un vocabulario de 27.971 tokens.

**Configuración del modelo 1:**

- Neuronas ocultas: 130
- Tamaño de contexto: 2
- Épocas: 1250
- Tasa de aprendizaje: 0.001
- Palabras mal predichas: 3790

**Resultados finales del modelo 1:**

| Métrica | Valor |
|---|---|
| Top1 | 67.72 % |
| Top5 | 86.32 % |

**Configuración del modelo 2:**

- Neuronas ocultas: 100
- Tamaño de contexto: 4
- Épocas: 1600
- Tasa de aprendizaje: 0.001

**Resultados finales del modelo 2:**

| Métrica | Valor |
|---|---|
| Top1 | 65.79 % |
| Top5 | 86.01 % |

# Palabras que no se predijeron bien

| | | | | | |
|---|---|---|---|---|---|
| . | 2818 | vos | 119 | parece | 55 |
| , | 2242 | las | 118 | todavía | 54 |
| y | 1552 | ahora | 117 | cosas | 54 |
| . | 1316 | te | 117 | viejo | 54 |
| que | 1085 | talita | 112 | va | 53 |
| de | 1068 | maga | 111 | esto | 53 |
| no | 782 | mismo | 108 | gregorovius | 53 |
| a | 776 | sí | 102 | ver | 52 |
| en | 669 | horacio | 99 | día | 51 |
| pero | 576 | poco | 99 | tarde | 51 |
| lo | 479 | está | 98 | hombre | 50 |
| la | 478 | ahí | 97 | solamente | 49 |
| el | 430 | había | 95 | realidad | 48 |
| con | 417 | otro | 94 | hace | 48 |
| se | 384 | bien | 91 | mundo | 48 |
| como | 373 | noche | 90 | tanto | 47 |
| es | 357 | menos | 90 | mientras | 47 |
| por | 350 | traveler | 90 | sé | 47 |
| o | 308 | ella | 86 | entre | 45 |
| ? | 299 | vez | 86 | rocamadour | 45 |
| para | 299 | sin | 85 | mucho | 45 |
| era | 292 | mejor | 85 | dice | 44 |
| un | 282 | dos | 84 | fuera | 44 |
| una | 281 | entonces | 84 | cama | 44 |
| yo | 247 | casi | 83 | boca | 44 |
| cuando | 244 | hay | 82 | etienne | 44 |
| más | 240 | uno | 81 | fue | 43 |
| oliveira | 238 | tiempo | 80 | puerta | 43 |
| eso | 235 | también | 79 | verdad | 42 |
| si | 234 | otra | 74 | babs | 42 |
| le | 216 | donde | 72 | veces | 42 |
| ya | 212 | mano | 69 | desde | 42 |
| dijo | 205 | tenía | 69 | antes | 41 |
| todo | 203 | hacer | 68 | claro | 41 |
| me | 191 | tan | 68 | pensar | 40 |
| porque | 181 | ese | 67 | lado | 40 |
| del | 179 | vida | 66 | mal | 40 |
| los | 171 | aunque | 65 | mí | 40 |
| al | 165 | tiene | 63 | ventana | 39 |
| así | 148 | casa | 63 | aire | 39 |
| ; | 144 | sobre | 62 | pieza | 39 |
| qué | 142 | ser | 62 | sueño | 39 |
| su | 140 | usted | 61 | ni | 39 |
| : | 129 | esa | 61 | fin | 38 |
| nada | 123 | aquí | 60 | nos | 38 |
| él | 123 | decir | 59 | mañana | 38 |
| estaba | 123 | ronald | 57 | mirando | 38 |
| siempre | 122 | cosa | 57 | manú | 38 |
| hasta | 122 | mi | 57 | | |
| después | 122 | mamá | 57 | | |
| algo | 121 | cara | 55 | | |

# Análisis de contextos

Considerando los resultados obtenidos y tras realizar pruebas con distintos modelos que presentaron valores similares de acierto y error, se decidió analizar las frecuencias de las palabras que fueron incorrectamente predichas. A continuación, se muestran aquellas palabras que se predicen de forma errónea en más de 500 ocasiones, junto con las palabras que con mayor frecuencia aparecen a su derecha e izquierda. El objetivo de este análisis es combinar esta información posteriormente con el fin de mejorar el desempeño general del modelo.

**Palabras más frecuentes cerca de `'.'` (2818 apariciones):**

```
.: 1910
no: 421
pero: 324
la: 324
y: 318
```

**Palabras más frecuentes cerca de `','` (2242 apariciones):**

```
y: 1992
pero: 755
la: 696
que: 609
el: 576
```

**Palabras más frecuentes cerca de `'y'` (1552 apariciones):**

```
,: 1992
el: 411
la: 383
se: 323
.: 318
```

**Palabras más frecuentes cerca de `'. '` (1316 apariciones):**

```
no: 227
.: 204
oliveira: 116
la: 113
por: 112
```

**Palabras más frecuentes cerca de `'que'` (1085 apariciones):**

```
lo: 788
,: 609
se: 571
no: 503
de: 475
```

**Palabras más frecuentes cerca de `'de'` (1068 apariciones):**

```
la: 1854
los: 659
,: 484
que: 475
las: 466
```

**Palabras más frecuentes cerca de `'no'` (782 apariciones):**

```
,: 553
que: 503
.: 421
se: 364
. : 227
```

**Palabras más frecuentes cerca de `'a'` (776 apariciones):**

```
la: 977
,: 417
.: 253
y: 249
los: 245
```

**Palabras más frecuentes cerca de `'en'` (669 apariciones):**

```
el: 1219
la: 1101
,: 435
que: 394
un: 336
```

**Palabras más frecuentes cerca de 'pero' (576 apariciones):**

```
,: 755
.: 324
no: 145
. : 91
el: 58
```

## Corpus Fragmentado por Frecuencias

**Características del corpus:** Se utilizó un modelo de fragmentación explicado anteriormente, dando un total de 289.919 palabras en el corpus y un vocabulario de 28.030 tokens. Se agruparon en un único *token* aquellas palabras que aparecían más de 500 veces y que, además, compartían un contexto inmediato (de una posición de distancia) en más de 200 ocasiones. Este proceso se aplicó de forma iterativa hasta que no fue posible realizar más combinaciones (*merges*).

**Configuración del modelo:**

- Neuronas ocultas: 130
- Tamaño de contexto: 5
- Épocas: 1000
- Tasa de aprendizaje: 0.01
- Palabras mal predichas: 132

**Resultados finales:**

| Métrica | Valor |
|---|---|
| Top1 | 94.50 % |
| Top5 | 99.10 % |

# Palabras que no se predijeron bien

| | | | | | |
|---|---|---|---|---|---|
| , | 450 | sí | 7 | entonces | 1 |
| . | 321 | gregorovius | 7 | hacer | 1 |
| . | 307 | lo que | 6 | , en | 1 |
| de | 127 | él | 6 | no , | 1 |
| y | 109 | a la | 6 | fuera | 1 |
| que | 105 | siempre | 6 | de que | 1 |
| a | 101 | los | 5 | hombre | 1 |
| no | 92 | , no | 5 | aquí | 1 |
| la | 68 | decir | 5 | sabe | 1 |
| se | 55 | mi | 4 | casi | 1 |
| en | 47 | bien | 4 | que no | 1 |
| el | 45 | porque | 4 | mucho | 1 |
| por | 41 | sin | 4 | ossip | 1 |
| es | 37 | después | 4 | wong | 1 |
| me | 36 | ella | 4 | empezó | 1 |
| , y | 36 | otro | 4 | suerte | 1 |
| lo | 36 | qué | 3 | desde | 1 |
| talita | 35 | cuando | 3 | etienne | 1 |
| con | 32 | mejor | 3 | . no | 1 |
| ? | 27 | estaba | 3 | usted | 1 |
| un | 26 | babs | 3 | miedo | 1 |
| yo | 25 | menos | 3 | viejo | 1 |
| era | 17 | tiempo | 3 | acuerdo | 1 |
| eso | 17 | , que | 3 | cosas | 1 |
| o | 15 | mismo | 3 | entrar | 1 |
| del | 14 | ahora | 2 | apenas | 1 |
| una | 14 | : | 2 | maga | 1 |
| todo | 14 | dos | 2 | quiero | 1 |
| oliveira | 13 | tarde | 2 | manú | 1 |
| como | 13 | vez | 2 | quiere | 1 |
| para | 13 | así | 2 | donde | 1 |
| le | 11 | su | 2 | , de | 1 |
| horacio | 11 | pero | 2 | pensar | 1 |
| más | 11 | había | 2 | parece | 1 |
| te | 11 | otra | 2 | izquierda | 1 |
| ya | 10 | en el | 2 | estar | 1 |
| dijo | 10 | las | 2 | calor | 1 |
| traveler | 10 | hay | 2 | todavía | 1 |
| . . | 9 | algo | 2 | poco | 1 |
| vos | 9 | dijo oliveira | 2 | esto | 1 |
| si | 8 | va | 2 | ahí | 1 |
| , pero | 8 | sé | 2 | aunque | 1 |
| al | 8 | está | 2 | | |
| . . . | 7 | ronald | 2 | | |
| de la | 7 | hablar | 1 | | |

# Análisis de contextos

**Palabras más frecuentes cerca de , (450 apariciones):**

```
una: 279
los: 265
por: 241
es: 232
como: 222
```

**Palabras más frecuentes cerca de . (321 apariciones):**

```
no: 201
. .: 190
la: 112
por: 112
y: 97
```

**Palabras más frecuentes cerca de . (307 apariciones):**

```
a: 224
en: 192
por: 169
es: 143
se: 141
```

**Palabras más frecuentes cerca de de (127 apariciones):**

```
su: 194
y: 149
después: 107
.: 102
lo: 100
```

**Palabras más frecuentes cerca de y (109 apariciones):**

```
la: 285
se: 272
los: 188
a: 159
de: 149
```

**Palabras más frecuentes cerca de que (105 apariciones):**

```
en: 247
me: 195
es: 183
le: 179
hay: 178
```

**Palabras más frecuentes cerca de a (101 apariciones):**

```
los: 235
.: 224
va: 209
un: 195
las: 191
```

**Palabras más frecuentes cerca de no (92 apariciones):**

```
se: 262
. : 201
es: 133
ya: 122
me: 115
```

**Palabras más frecuentes cerca de la (68 apariciones):**

```
y: 285
maga: 271
por: 266
con: 265
dijo: 146
```

**Palabras más frecuentes cerca de se (55 apariciones):**

```
y: 272
no: 262
.: 141
le: 109
había: 104
```

**Palabras más frecuentes cerca de en (47 apariciones):**

```
que: 247
los: 229
```

```
una: 225
.: 192
las: 177
```

**Palabras más frecuentes cerca de `el` (45 apariciones):**

```
por: 232
con: 223
todo: 119
, y: 96
. : 91
```

**Palabras más frecuentes cerca de `por` (41 apariciones):**

```
la: 266
qué: 262
,: 241
el: 232
.: 169
```

**Palabras más frecuentes cerca de `es` (37 apariciones):**

```
,: 232
que: 183
.: 143
no: 133
un: 131
```

**Palabras más frecuentes cerca de `me` (36 apariciones):**

```
que: 195
,: 167
.: 126
no: 115
y: 80
```

**Palabras más frecuentes cerca de `, y` (36 apariciones):**

```
el: 96
la: 86
que: 76
en: 62
después: 59
```

**Palabras más frecuentes cerca de `lo` (36 apariciones):**

```
a: 162
mejor: 157
,: 153
por: 118
mismo: 105
```

**Palabras más frecuentes cerca de `talita` (35 apariciones):**

```
dijo: 92
.: 89
,: 73
. : 48
a: 37
```

**Palabras más frecuentes cerca de `con` (32 apariciones):**

```
la: 265
el: 223
un: 201
una: 140
los: 123
```

**Palabras más frecuentes cerca de `?` (27 apariciones):**

```
no: 69
qué: 40
dijo: 29
,: 25
preguntó: 24
```

**Palabras más frecuentes cerca de `un` (26 apariciones):**

```
poco: 277
como: 224
con: 201
a: 195
es: 131
```

**Palabras más frecuentes cerca de `yo` (25 apariciones):**

```
,: 171
```

```
.: 141
que: 90
y: 78
no: 72
```

**Palabras más frecuentes cerca de `era` (17 apariciones):**

```
,: 110
que: 88
no: 79
un: 77
una: 72
```

**Palabras más frecuentes cerca de `eso` (17 apariciones):**

```
,: 114
que: 96
todo: 79
por: 70
es: 62
```

**Palabras más frecuentes cerca de `o` (15 apariciones):**

```
,: 194
de: 58
el: 38
dos: 37
en: 35
```

**Palabras más frecuentes cerca de `del` (14 apariciones):**

```
,: 52
mundo: 47
lado: 46
colorado: 39
tiempo: 35
```

**Palabras más frecuentes cerca de `una` (14 apariciones):**

```
,: 279
en: 225
con: 140
como: 134
es: 118
```

**Palabras más frecuentes cerca de `todo` (14 apariciones):**

```
el: 119
,: 91
eso: 79
.: 78
de: 74
```

**Palabras más frecuentes cerca de `oliveira` (13 apariciones):**

```
,: 115
.: 113
a: 76
se: 70
. : 66
```

**Palabras más frecuentes cerca de `como` (13 apariciones):**

```
un: 224
,: 222
si: 211
una: 134
es: 84
```

**Palabras más frecuentes cerca de `para` (13 apariciones):**

```
que: 147
,: 101
.: 64
la: 47
no: 44
```

**Palabras más frecuentes cerca de `le` (11 apariciones):**

```
que: 179
,: 131
se: 109
y: 92
no: 91
```

**Palabras más frecuentes cerca de `horacio` (11 apariciones):**

```
,: 102
. : 38
.: 29
a: 23
se: 20
```

**Palabras más frecuentes cerca de `más` (11 apariciones):**

```
que: 161
,: 71
vez: 67
```

```
lo: 59
bien: 59
```

**Palabras más frecuentes cerca de `te` (11 apariciones):**

```
que: 100
no: 79
.: 61
,: 58
vos: 38
```

## Corpus Fragmentado con BPE

**Características del corpus:** Se utilizó un modelo de fragmentación BPE con 10.000 merges, dando un total de 359.142 palabras en el corpus y un vocabulario de 9.676 tokens.

**Configuración del modelo:**

- Neuronas ocultas: 130
- Tamaño de contexto: 15
- Épocas: 2000
- Tasa de aprendizaje: 0.01
- Palabras mal predichas: 136

**Resultados finales:**

| Métrica | Valor |
|---|---|
| Top1 | 81.10 % |
| Top5 | 95.53 % |

# Análisis de contextos

**Palabras más frecuentes cerca de .</w> (4209 apariciones):**

```
.</w>: 2320
no</w>: 651
la</w>: 445
y</w>: 430
pero</w>: 415
```

**Palabras más frecuentes cerca de ,</w> (3738 apariciones):**

```
y</w>: 2016
pero</w>: 757
la</w>: 720
que</w>: 619
el</w>: 585
```

**Palabras más frecuentes cerca de el</w> (1077 apariciones):**

```
en</w>: 1224
,</w>: 585
que</w>: 477
y</w>: 415
.</w>: 408
```

**Palabras más frecuentes cerca de de</w> (999 apariciones):**

```
la</w>: 1867
los</w>: 670
,</w>: 505
las</w>: 483
que</w>: 481
```

**Palabras más frecuentes cerca de a</w> (857 apariciones):**

```
la</w>: 999
,</w>: 458
.</w>: 357
y</w>: 263
los</w>: 251
```

**Palabras más frecuentes cerca de la</w> (738 apariciones):**

```
de</w>: 1867
en</w>: 1115
a</w>: 999
,</w>: 720
maga</w>: 463
```

**Palabras más frecuentes cerca de y</w> (651 apariciones):**

```
,</w>: 2016
.</w>: 430
el</w>: 415
la</w>: 393
se</w>: 326
```

**Palabras más frecuentes cerca de que</w> (642 apariciones):**

```
lo</w>: 790
,</w>: 619
se</w>: 573
no</w>: 504
de</w>: 481
```

**Palabras más frecuentes cerca de en</w> (587 apariciones):**

```
el</w>: 1224
la</w>: 1115
,</w>: 459
que</w>: 405
.</w>: 374
```

# Palabras que no se predijeron bien

| | | | | | |
|---|---|---|---|---|---|
| .</w> | 4209 | más</w> | 3 | frío</w> | 1 |
| ,</w> | 3738 | algo</w> | 3 | mucho</w> | 1 |
| el</w> | 1077 | está</w> | 3 | siquiera</w> | 1 |
| de</w> | 999 | había</w> | 3 | tablón</w> | 1 |
| a</w> | 857 | les</w> | 3 | su</w> | 1 |
| la</w> | 738 | gregorovius</w> | 3 | dice</w> | 1 |
| y</w> | 651 | maga</w> | 2 | esperá</w> | 1 |
| que</w> | 642 | babs</w> | 2 | ser</w> | 1 |
| en</w> | 587 | ese</w> | 2 | claro</w> | 1 |
| un</w> | 388 | gran</w> | 2 | cuenta</w> | 1 |
| no</w> | 385 | camas</w> | 2 | kibbutz</w> | 1 |
| se</w> | 354 | cuando</w> | 2 | sopa</w> | 1 |
| oliveira</w> | 266 | noche</w> | 2 | bueno</w> | 1 |
| del</w> | 252 | amor</w> | 2 | otro</w> | 1 |
| lo</w> | 196 | qué</w> | 2 | tan</w> | 1 |
| dijo</w> | 123 | cosas</w> | 2 | aunque</w> | 1 |
| por</w> | 73 | tiempo</w> | 2 | e</w> | 1 |
| como</w> | 66 | decía</w> | 2 | ahí</w> | 1 |
| una</w> | 48 | lucía</w> | 1 | ventana</w> | 1 |
| los</w> | 42 | yes</w> | 1 | dos</w> | 1 |
| me</w> | 31 | re | 1 | eleg | 1 |
| ?</w> | 28 | porque</w> | 1 | ke</w> | 1 |
| con</w> | 24 | oh</w> | 1 | apenas</w> | 1 |
| le</w> | 16 | ya</w> | 1 | hace</w> | 1 |
| te</w> | 14 | entonces</w> | 1 | malo</w> | 1 |
| era</w> | 13 | estás</w> | 1 | pieza</w> | 1 |
| es</w> | 12 | ver | 1 | puerta</w> | 1 |
| las</w> | 9 | sin</w> | 1 | cama<</w> | 1 |
| traveler</w> | 9 | esta</w> | 1 | todavía</w> | 1 |
| yo</w> | 9 | siempre</w> | 1 | i | 1 |
| horacio</w> | 8 | club</w> | 1 | rojo</w> | 1 |
| así</w> | 6 | sé</w> | 1 | amarillo</w> | 1 |
| todo</w> | 6 | just | 1 | horrible</w> | 1 |
| tu</w> | 6 | muy</w> | 1 | efectos/w> | 1 |
| hasta</w> | 5 | bajo</w> | 1 | estaba</w> | 1 |
| eso</w> | 5 | enormemente</w> | 1 | hacía</w> | 1 |
| vos</w> | 4 | ossip</w> | 1 | ver</w> | 1 |
| ella</w> | 4 | nos</w> | 1 | ojos</w> | 1 |
| sí</w> | 4 | aba</w> | 1 | papeles</w> | 1 |
| ronald</w> | 4 | célestin</w> | 1 | entrada</w> | 1 |
| al</w> | 4 | podían</w> | 1 | s | 1 |
| o</w> | 4 | tanto</w> | 1 | para</w> | 1 |
| gekrepten</w> | 4 | señora</w> | 1 | vida</w> | 1 |
| os</w> | 3 | jo | 1 | café</w> | 1 |
| después</w> | 3 | tarde</w> | 1 | | |
| pero</w> | 3 | cualquier</w> | 1 | | |

A continuación, se presentan los mejores valores obtenidos para cada método evaluado, considerando las métricas **Top1** y **Top5**. Estos resultados permiten comparar el rendimiento general de los distintos métodos en términos de precisión de predicción de la palabra central.

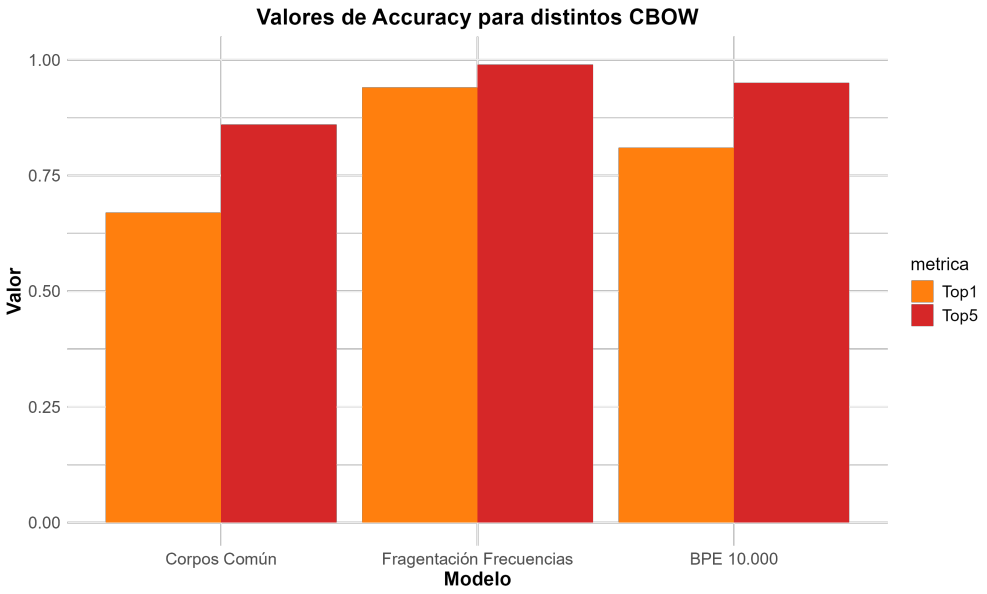


Figura 1: Comparación de las métricas **Top1** y **Top5** para los distintos métodos evaluados.

Como se observa en la Figura 1, los métodos fragmentados presentan comportamientos similares en **Top5**, pero sin dudas hay un claro ganados. El análisis de estas variaciones permite seleccionar los parámetros más adecuados y determinar qué configuraciones de entrenamiento logran un mejor equilibrio entre exactitud y generalización. Como prueba quedaría ver si continuando las agrupaciones de tokens con umbrales más chicos mejora el método.

!! Muchas gracias ¡¡