# Estudio de diferentes modelos de redes neuronales para el desarrollo de un clasificador de frases

Arkaitz Bidaurrazaga Barrueta

María Inés Torres
Departamento de Electricidad y Electrónica (EHU/UPV)

Raquel Justo Blanco
Departamento de Electricidad y Electrónica (EHU/UPV)

Julio 2019

# Introducción

- Problema : Clasificar frases

# Introducción

- Problema : Clasificar frases
- Objetivo : Desarrollar una red neuronal que haga el proceso de NLU

# Introducción

- Problema : Clasificar frases
- Objetivo : Desarrollar una red neuronal que haga el proceso de NLU
- Metodología : Aprendizaje supervisado

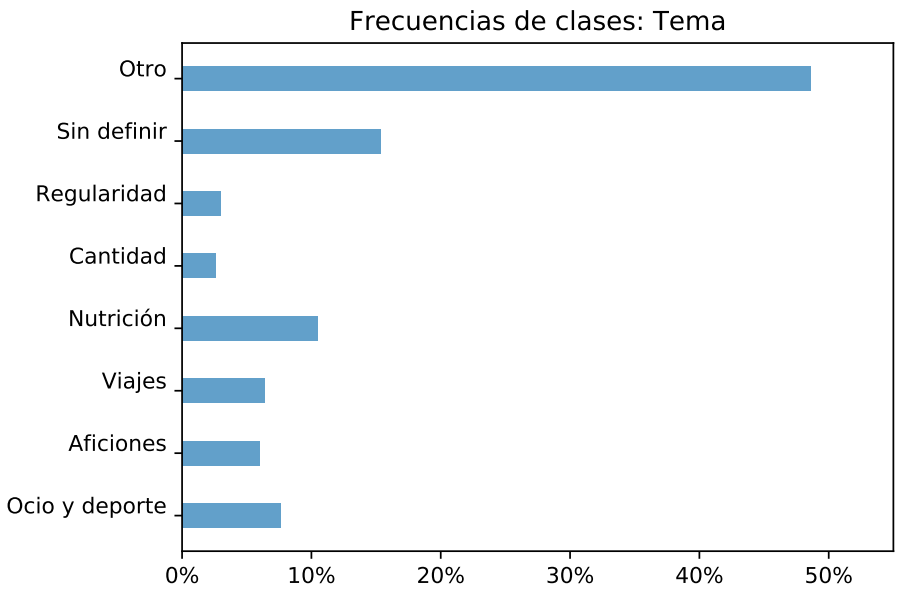
# Tabla de contenidos

# Tabla de contenidos

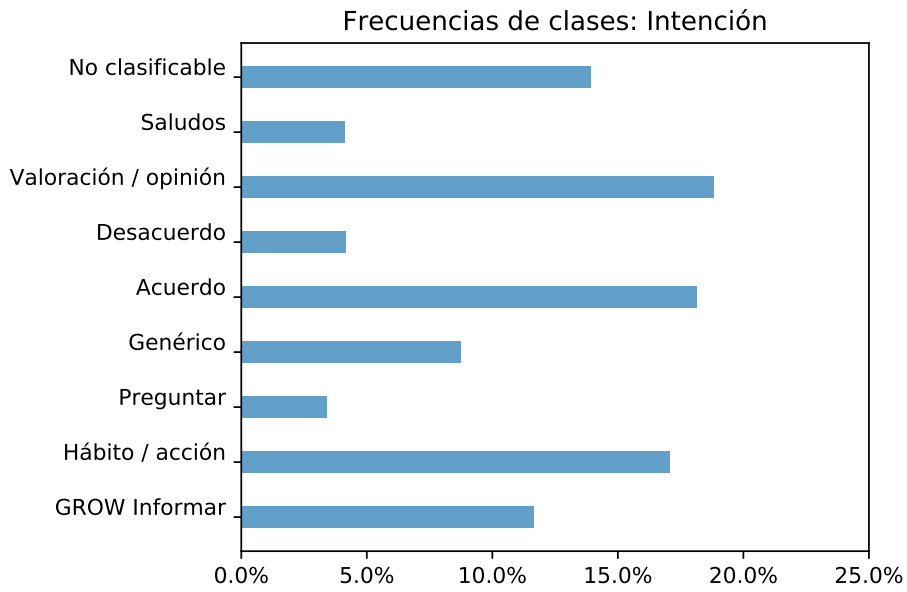
# Corpus de EMPATHIC

| | |
|---|---|
| COACH: | Y entonces, para priorizar eso, ¿qué podrías hacer? |
| USUARIO: | Pues hombre, se me ocurre que podría hacer ejercicio más a menudo y salir a más conciertos de los que me gustan. |
| POLARIDAD: | Neutra. |
| SUBFRASE 1: | podría hacer ejercicio más a menudo |
| ETIQUETA TEMA: | tema, ocio y deporte, deportes, frecuencia |
| ETIQUETA INTENCIÓN: | intención, informar, plan, posible/no definitivo |
| ENTIDADES: | cantidades: más; fechas relativas: a menudo; acción: hacer ejercicio |
| SUBFRASE 2: | salir a más conciertos de los que me gustan |
| ETIQUETA TEMA: | tema, ocio y deporte, eventos, espectador |
| ETIQUETA INTENCIÓN: | intención, informar, plan, posible/no definitivo |
| ENTIDADES: | cantidades: más; afición: conciertos |

# Frecuencias de clases




(a) Tema




(b) Intención

# Tamaño del corpus

- Número de clases por ámbito:

| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

# Tamaño del corpus

- Número de clases por ámbito:

| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

- Diálogos : 140

# Tamaño del corpus

- Número de clases por ámbito:

| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

- Diálogos : 140
- Subfrases : 7500

# Tamaño del corpus

- Número de clases por ámbito:

| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

- Diálogos : 140
- Subfrases : 7500
- Vocabulario completo : 4400 palabras

# Tamaño del corpus

- Número de clases por ámbito:

| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

- Diálogos : 140
- Subfrases : 7500
- Vocabulario completo : 4400 palabras
- Vocabulario *wordvector* : 2300

# Tamaño del corpus

- Número de clases por ámbito:

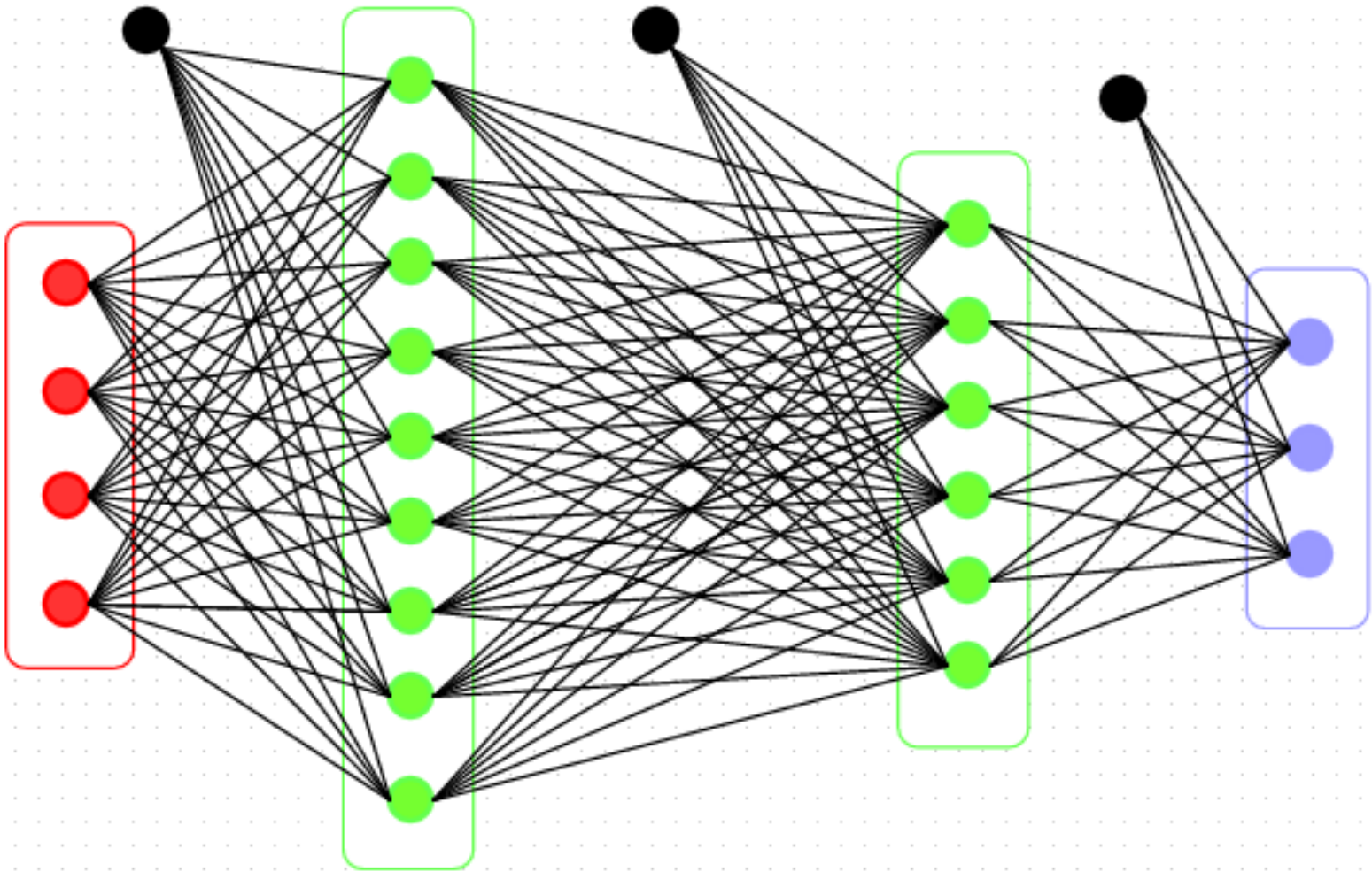
| | Tema | Intención | Polaridad |
|---|---|---|---|
| $M$ | 8 | 9 | 3 |

- Diálogos : 140
- Subfrases : 7500
- Vocabulario completo : 4400 palabras
- Vocabulario *wordvector* : 2300
- Dimensión de los *wordvectors* : 300

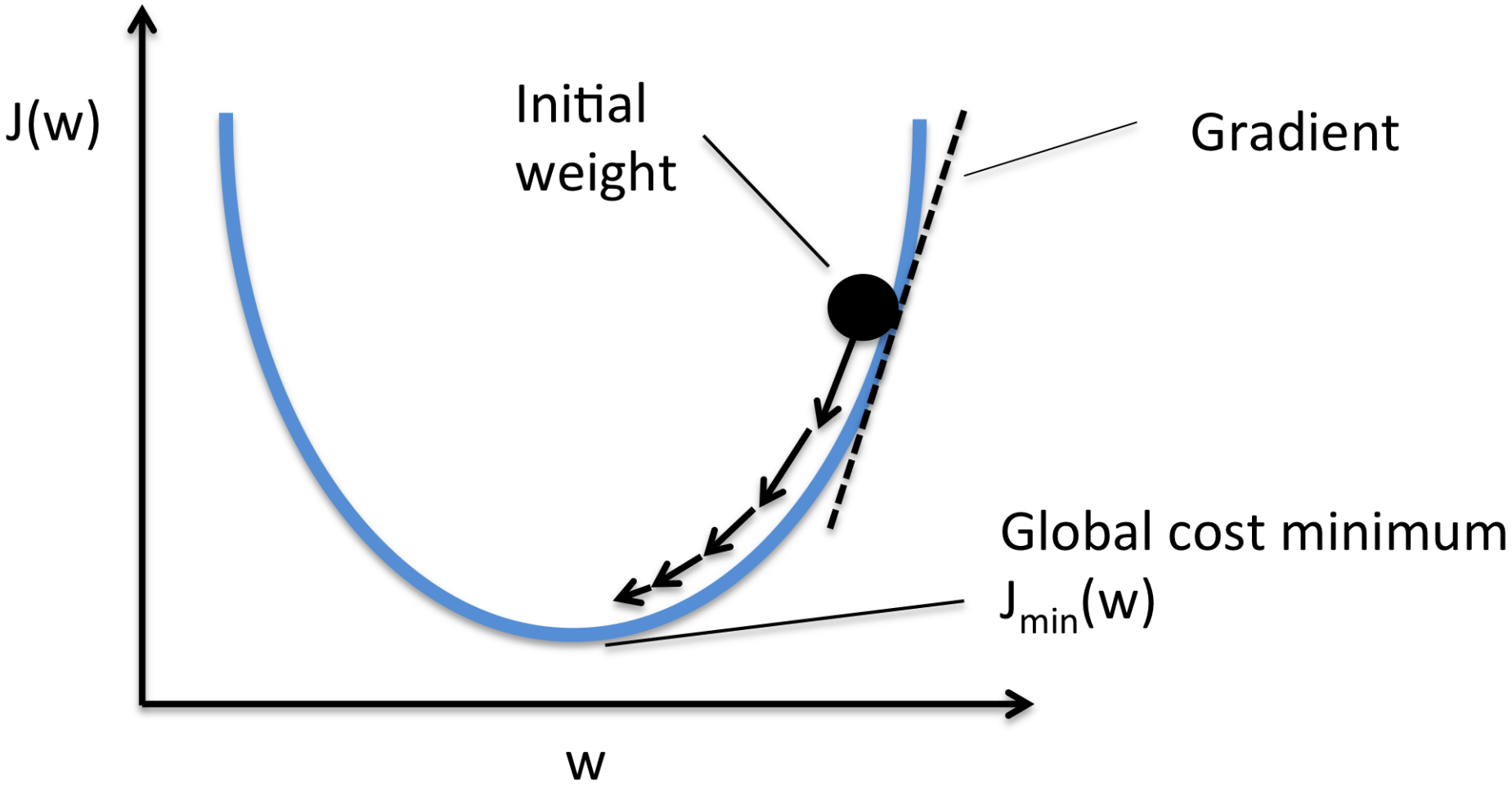
# Tabla de contenidos

# ¿Qué es una red neuronal?

$$w_i, \beta_j \equiv \theta$$

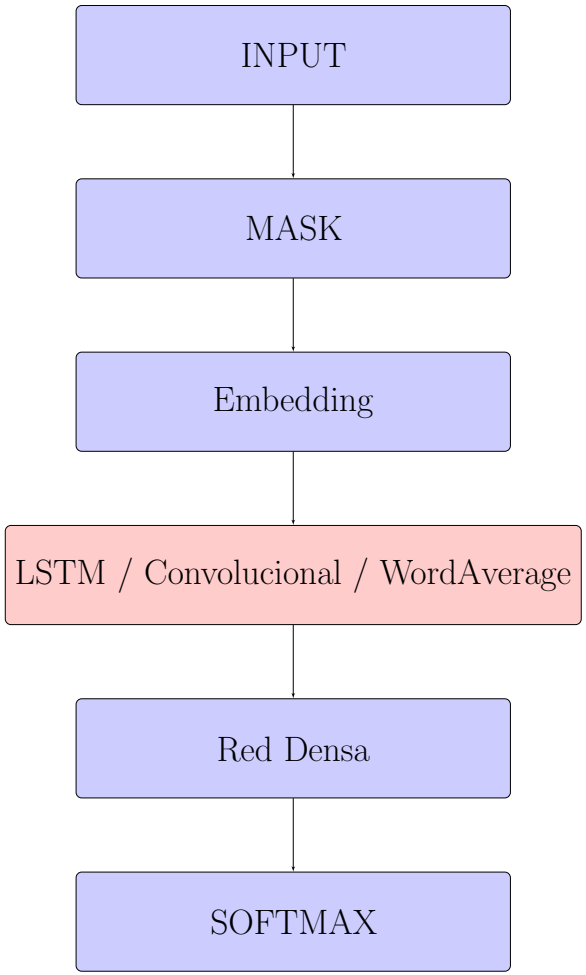


# Descenso por gradiente



# Arquitectura y datos de entrada

Arquitectura:




Datos de entrada:

- Frase en el corpus :
  "Hola PERSON . ¿ Qué tal estás ?"
- Secuencia de índices :
  [ 109, 38, 1, 32, 73, 336, 270, 31 ]

# Tabla de contenidos

# Gradiente del error

Siendo $\boldsymbol{g}_t$ la gradiente de la función de error $L$ respecto a los parámetros $\boldsymbol{\theta}_t$, en cada momento del entrenamiento $t$:

$$\boldsymbol{g}_t = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}_t} L(\mathrm{x}^i, y^i, \boldsymbol{\theta}_t) \tag{1}$$

# Adam

Actualización de los parámetros:

$$
\begin{aligned}
\mathsf{s}_t &= \rho_1 \mathsf{s}_{t-1} + (1-\rho_1)\boldsymbol{g}_t \\
\mathsf{r}_t &= \rho_2 \mathsf{r}_{t-1} + (1-\rho_2)\boldsymbol{g}_t^2 \\
\hat{\mathsf{s}}_t &= \frac{\mathsf{s}_t}{1-\rho_1^t} \\
\hat{\mathsf{r}}_t &= \frac{\mathsf{r}_t}{1-\rho_2^t} \\
\Delta\boldsymbol{\theta}_t &= -\epsilon \frac{\hat{\mathsf{s}}_t}{\delta + \sqrt{\hat{\mathsf{r}}_t}} \\
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}_t
\end{aligned}
\tag{2}
$$

Hiperparámetros:

- $\epsilon = 0.001$
- $\rho_1 = 0.9$
- $\rho_2 = 0.999$
- $\delta = 10^{-8}$

# *Nesterov momentum*

Actualización de los parámetros:

$$\boldsymbol{g}_t = \frac{1}{m'} \sum_{i=1}^{m'} \nabla_{\boldsymbol{\theta}_t} L(\mathrm{x}^i, y^i, \boldsymbol{\theta}_t + \alpha \boldsymbol{v}_t)$$

$$\boldsymbol{v}_{t+1} = \alpha \boldsymbol{v}_t - \epsilon \boldsymbol{g}_t$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{v}_{t+1}$$

Hiperparámetros:

- $\epsilon$ : ratio de aprendizaje
- $\alpha$ : parámetro de momento

# Regularizadores

Penalización de parámetros (pesos $w_i$ y umbrales $\beta_j$) demasiado "grandes":

$$\begin{aligned} L1: \quad & L'(x^i, y^i) = L(x^i, y^i) + \lambda \sum_i |w_i| + \mu \sum_j |\beta_j| \\ L2: \quad & L'(x^i, y^i) = L(x^i, y^i) + \lambda \sum_i w_i^2 + \mu \sum_j \beta_j^2 \end{aligned} \tag{3}$$

Función de error:

$$L(x, y) = -\sum_{i=1}^{M} y_i \log(x_i) \tag{4}$$

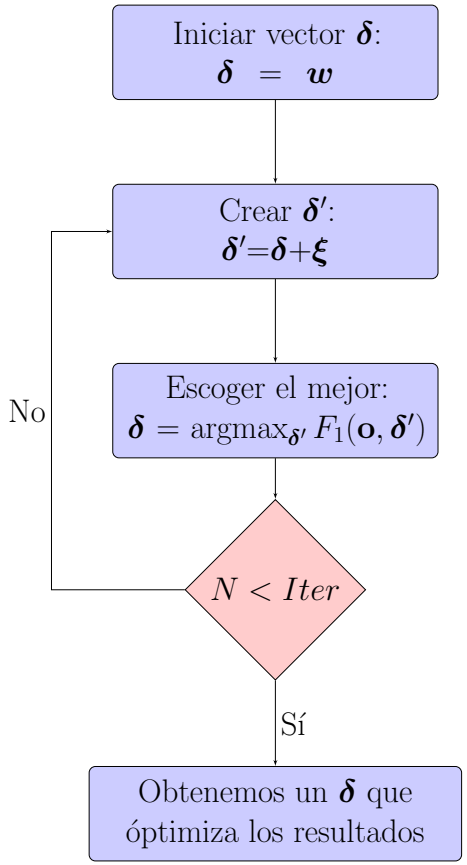
# Tabla de contenidos

# F1 *Score* y Búsqueda local

- F1 *Score* más apropiado que la Exactitud

# F1 *Score* y Búsqueda local

- F1 *Score* más apropiado que la Exactitud
- Uso de algoritmo de Búsqueda local



# Tabla de contenidos

# Rendimiento de los clasificadores

| | **Intención** | | |
|---|---|---|---|
| | *LSTM* | *Convolucional* | *WordAverage* |
| $E$ | 37 % | 48 % | 57 % |
| $E(\boldsymbol{\delta})$ | 39 % | 46 % | 57 % |
| $F_1$ | 20 % | 50 % | **60** % |
| $F_1(\boldsymbol{\delta})$ | 27 % | 52 % | **59** % |

| | **Tema** | | |
|---|---|---|---|
| | *LSTM* | *Convolucional* | *WordAverage* |
| $E$ | 50 % | 62 % | 66 % |
| $E(\boldsymbol{\delta})$ | 49 % | 53 % | 66 % |
| $F_1$ | 10 % | 23 % | 42 % |
| $F_1(\boldsymbol{\delta})$ | 26 % | 28 % | 45 % |

| | **Polaridad** | | |
|---|---|---|---|
| | *LSTM* | *Convolucional* | *WordAverage* |
| $E$ | 64 % | 62 % | 64 % |
| $E(\boldsymbol{\delta})$ | 28 % | 54 % | 60 % |
| $F_1$ | 26 % | 36 % | 40 % |
| $F_1(\boldsymbol{\delta})$ | 19 % | 41 % | 40 % |

# Optimización de la red

*Notación: $L1(\lambda, \mu)$ , $L2(\lambda, \mu)$ , $N(\epsilon, \alpha)$

| | ***WordAverage*: Intención** | | | | | |
|---|---|---|---|---|---|---|
| | $L1(10^{-2}, 10^{-2})$ | $L1(0, 10^{-3})$ | $L2(10^{-4}, 10^{-4})$ | $N(10^{-1}, 10^{-3})$ | $N(1, 10^{-3})$ | $N(1, 10^{-3})\|L2(10^{-4}, 10^{-4})$ |
| $E$ | 31 % | 59 % | 60 % | 53 % | **60 %** | 62 % |
| $E(\boldsymbol{\delta})$ | 35 % | 59 % | 59 % | 52 % | **61 %** | 62 % |
| $F_1$ | 11 % | 62 % | 64 % | 55 % | **63 %** | 65 % |
| $F_1(\boldsymbol{\delta})$ | 28 % | 62 % | 63 % | 54 % | **65 %** | 64 % |

# Tabla de confusión con el mejor $F_1$ Score

| Real\Predicho | GROW Informar | Hábito / acción | Preguntar | Genérico | Acuerdo | Desacuerdo | Valoración / opinión | Saludos | No clasificable |
|---|---|---|---|---|---|---|---|---|---|
| GROW Informar | 45 | 54 | 0 | 7 | 2 | 1 | 38 | 1 | 13 |
| Hábito / acción | 15 | 173 | 0 | 4 | 5 | 1 | 18 | 1 | 26 |
| Preguntar | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 |
| Genérico | 10 | 11 | 4 | 58 | 4 | 1 | 24 | 0 | 26 |
| Acuerdo | 0 | 5 | 0 | 3 | 241 | 0 | 14 | 0 | 7 |
| Desacuerdo | 3 | 1 | 0 | 3 | 1 | 51 | 2 | 0 | 0 |
| Valoración / opinión | 27 | 37 | 1 | 11 | 21 | 3 | 177 | 3 | 21 |
| Saludos | 0 | 1 | 1 | 1 | 1 | 0 | 3 | 67 | 2 |
| No clasificable | 28 | 39 | 3 | 23 | 14 | 3 | 29 | 4 | 69 |

# Tabla de contenidos

# Conclusiones

1. Aprendido a trabajar con redes neuronales

# Conclusiones

1. Aprendido a trabajar con redes neuronales
2. Importancia del F1 *Score*

# Conclusiones

1. Aprendido a trabajar con redes neuronales
2. Importancia del F1 *Score*
3. Gran eficiencia con un modelo simple

# Conclusiones

1. Aprendido a trabajar con redes neuronales
2. Importancia del F1 *Score*
3. Gran eficiencia con un modelo simple
4. Importancia de un corpus balanceado

# Conclusiones

1. Aprendido a trabajar con redes neuronales
2. Importancia del F1 *Score*
3. Gran eficiencia con un modelo simple
4. Importancia de un corpus balanceado
5. Posibles mejoras en el futuro