UNIVERSIDAD TECNOLÓGICA NACIONAL

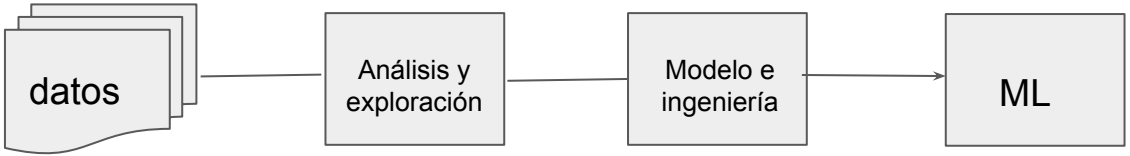
FACULTAD REGIONAL CÓRDOBA

Diplomatura en Data Science Aplicada

Agenda

- ML workflow
- Más clasificación de texto
- Análisis de sentimientos

Modelado un problema



- Qué datos tenemos, cuantos tenemos, faltan?
- Existen valores atípicos? qué hacemos con ellos?
- Ganamos conocimiento de negocio
- Planteamos hipótesis o supuestos
    - Respondemos a preguntas
- Definimos alcances y acciones.

- Qué forma deben tener nuestros datos?
- Tenemos suficientes datos? No, que hacemos entonces?
- Limitaciones con las que contamos, recursos, tiempo.
- Qué modelo utilizamos? es el único?
    - Evolucionamos nuestra POC o cambiamos de framework.
    - Construimos varios modelos conectados?

Modelado un problema - Herramientas

Almacenamiento

Transformación

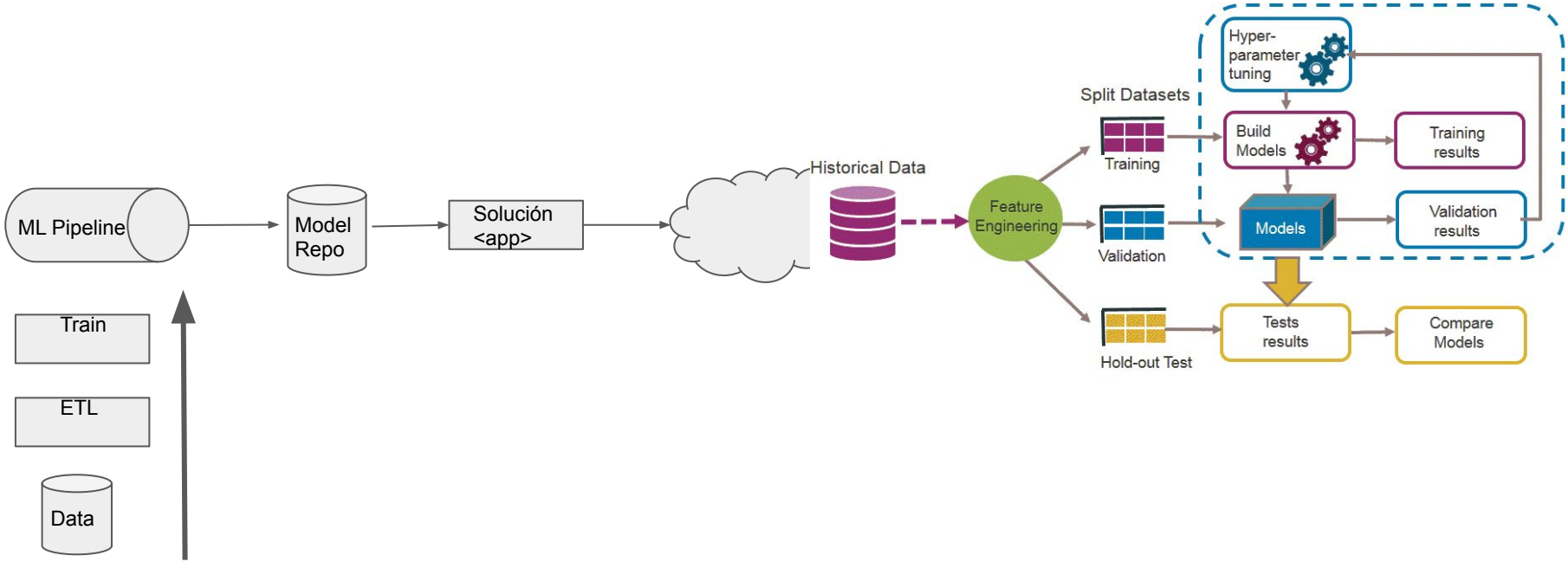
Visualización

Modelado

Monitoreo

Otros

Modelado un problema - Solución final



Text Classification

Volvamos

Text Classification - Problemáticas

Algunas aplicaciones comunes donde encontramos soluciones de clasificación de texto.

- Respuesta de preguntas automáticas (watson).
- Information Retrieval(duckduckgo).
- Extracción de información(Gmail structures events from emails).
- Machine Translation(Google Translate).
- Sentiment Analysis(Hater News).
- Text Summarization(Smmry o Reddit's autotldr).
- Detección de Spam Filter(Gmail).
- Speech Recognition(Apple Siri IA).

Text Classification - Definición formal

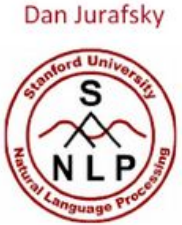
Dan Jurafsky

Text Classification: definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- *Output*: a predicted class $c \in C$

Text Classification - Reglas duras

Dan Jurafsky

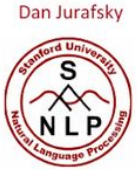


Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND“have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Text Classification - Modelos de ML

Dan Jurafsky



Classification Methods: Supervised Machine Learning

- *Input:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of m hand-labeled documents $(d_1, c_1), \ldots, (d_m, c_m)$
- *Output:*
  - a learned classifier $\gamma: d \rightarrow c$

Algunos Clasificadores:

- SVM
- KNN
- Logistic regression
- Naive Bayes

Text Classification - Tipos de clasificación

Dan Jurafsky

More Than Two Classes: Sets of binary classifiers

- One-of or multinomial classification
  - Classes are mutually exclusive: each document in exactly one class
- For each class $c \in C$
  - Build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test doc d,
  - Evaluate it for membership in each class using each γ_c
  - d belongs to the one class with maximum score

Text Classification - Tipos de clasificación

Dan Jurafsky

More Than Two Classes: Sets of binary classifiers

- Dealing with any-of or multivalue classification
  - A document can belong to 0, 1, or >1 classes.
- For each class $c \in C$
  - Build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test doc d,
  - Evaluate it for membership in each class using each γ_c
  - d belongs to any class for which γ_c returns true

Text Classification - Multinomial != Multivariante

Multivariante o Multivariable

Un modelo va tener más de una variable dependiente.Se trata de que tenemos 2 o más features que dependen uno de otros.

Multinomial
Estamos refiriéndonos a que las variables podrán tomar uno o más valores (tener más de 2 posibles valores finitos). Podemos pensarlo como la cantidad de categorías en la variable dependiente.

Entonces podemos tener un modelo que sea multivariante y multinomial.

Text Classification - Evaluación

Debemos poder medir que tan bien está clasificando nuestro modelo en las clases definidas.

Text Classification - Evaluación

Debemos poder medir que tan bien está clasificando nuestro modelo en las clases definidas.

→ True positive = correctly identified

→ False positive = incorrectly identified

→ True negative = correctly rejected

→ False negative = incorrectly rejected

Text Classification - Evaluando nuestro modelo

Dan Jurafsky

Per class evaluation measures

Recall:
Fraction of docs in class *i* classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:
Fraction of docs assigned class *i* that are actually about class *i*:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 - error rate)
Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

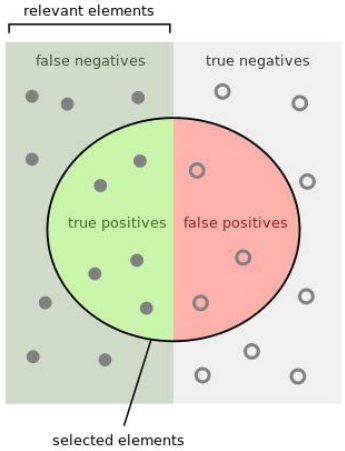
$$\text{Precision} = \frac{tp}{tp + fp}$$

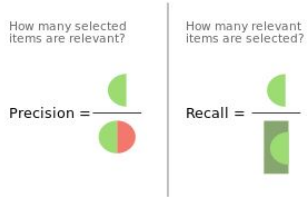
$$\text{Recall} = \frac{tp}{tp + fn}$$

accuracy (ACC)

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$





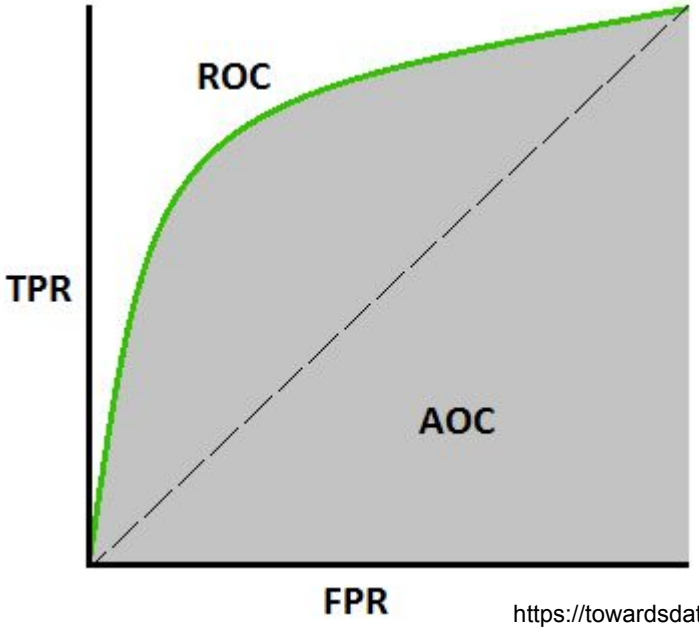
Un problema con F-Score es que no tiene en cuenta los true negatives.

https://en.wikipedia.org/wiki/Precision_and_recall

Text Classification - Evaluando nuestro modelo

"AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes."

Specificity
Son los **TNR,** y mide la proporción del valor actual de negativos que identificamos correctamente como tales.
→ FPR= 1 - specificity (TN/(TN+FP))



Consideraciones:
Cuando disminuimos el umbral, vamos a obtener un incremento en los valores positivos por lo que la sensibilidad aumenta.
En contrapartida, la especificidad del modelo va a disminuir.

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Text Classification - Evaluando múltiples clases

Dan Jurafsky

Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging**: Compute performance for each class, then average.
- **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.

Text Classification - Evaluando múltiples clases

Dan Jurafsky

Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83
- Microaveraged score is dominated by score on common classes

Text Classification - Analicemos un caso real

Por medio de un módulo de sklearn, podemos obtener algunas métricas.

→ *from sklearn.metrics import classification_report*

```
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Text Classification - Micro & Macro average ejemplo

	Gatos Predecidos	Perros Predecidos	Total
Gatos reales	80	20	100
Perros reales	50	50	100
Total	130	70	200

Supuesto:
Desarrollamos un clasificador que predice si una imágen hay una gato o un perro.

True positive: Un gato que fue identificado como gato.
False positive: Un perro que dijimos que es un gato.
True negative: Un perro que identificamos como perro.
False negative: Dijimos que era un gato cuando era perro

https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Text Classification - Set de datos

Dan Jurafsky

No training data? Manually written rules

If (wheat or grain) and not (whole or bread) then
Categorize as grain

- Need careful crafting
  - Human tuning on development data
  - Time-consuming: 2 days per class

Text Classification - Set de datos

Dan Jurafsky

Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a "high-bias" algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …

Text Classification - Set de datos

Dan Jurafsky

A reasonable amount of data?

- Perfect for all the clever classifiers
  - SVM
  - Regularized Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
  - Management likes quick fixes

Text Classification - Set de datos

Dan Jurafsky

A huge amount of data?

- Can achieve high accuracy!
- At a cost:
  - SVMs (train time) or kNN (test time) can be too slow
  - Regularized logistic regression can be somewhat better
- So Naïve Bayes can come back into its own again!

Text Classification - Salgan al sol

Es hora de arrancar a pensar un clasificador para el trabajo integrador

Text Classification

Sentiment Analysis

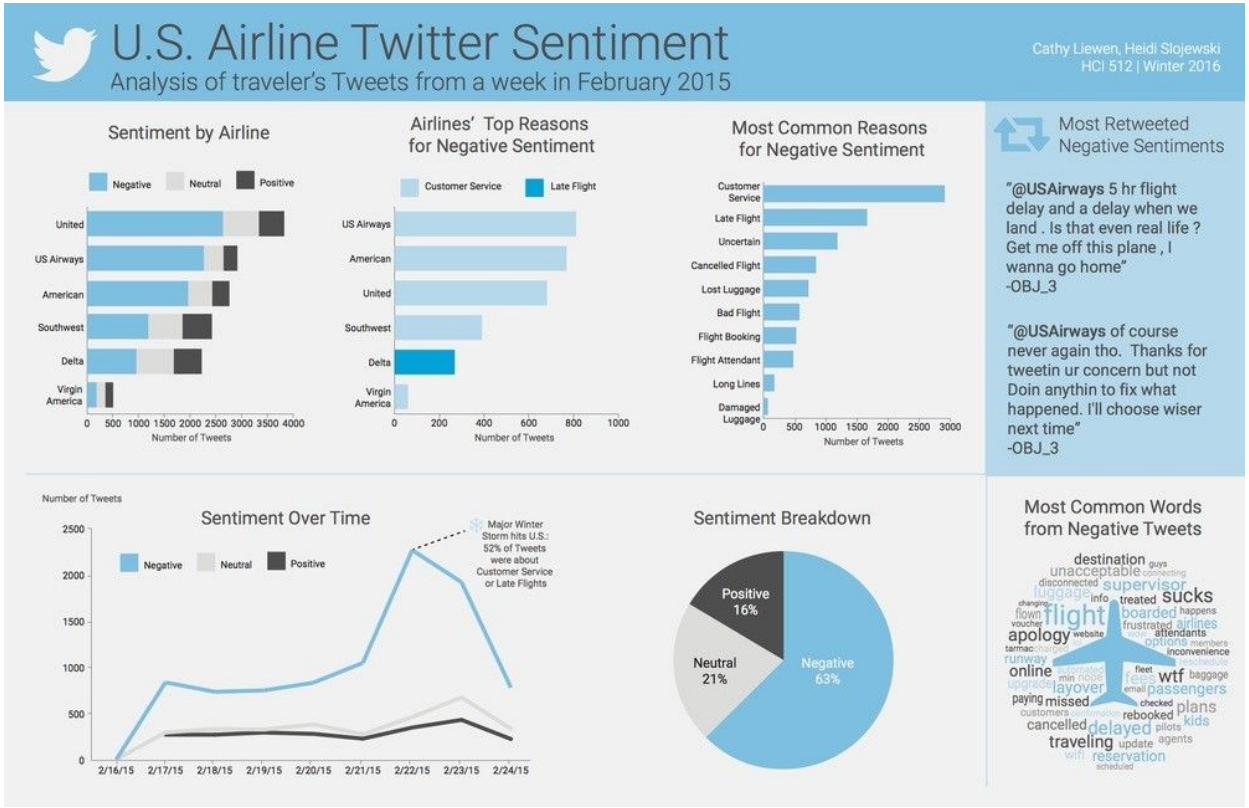
Sentiment Analysis

“Busca identificar y entender las emociones de las personas enfocado a un producto, suceso o evento.”

Tipos de valoración para una opinión:

- Positiva
- Negativa
- Neutra

Sentiment Analysis - Ejemplos



Sentiment Analysis - Ejemplos

Opiniones sobre el producto

4.9

Promedio entre 634 opiniones

5 estrellas	588
4 estrellas	40
3 estrellas	1
2 estrellas	1
1 estrella	4

Todas | Positivas | Negativas

La mejor calidad, los mejores resultados

Sin dudas la mejor. La estrené preparando una masa para pizza y fue todo lo que esperaba, un resultado excelente. La calidad del producto es inmejorable. Es un poco pesada pero es lógico porque requiere un motor potente. A diferencia del modelo chico, esta sí amasa. El modelo mini es mas liviano y chico pero no amasa. Es cara pero realmente lo vale. Todo esta calculado a la perfección. Incluso trae un circulo vertedor que te permite ir incorporando los ingredientes de costado a la vez que te protege por si la preparación llegara a salpicar. Y si sos fanático de la cocina, los accesorios que podes sumarle para hacer pastas, helados, picar, moler café, exprimir, etc, son increíbles. Más de 2 años

Denunciar

34 14

Muy bueno

Muy bueno sirve para muchas cosas con los accesorios que vienen para anexarles. Hace 2 meses

1 0

Excelente

Excelente la batidora, sin palabras. Funciona perfectamente. Hace 3 meses

0 0

Opiniones sobre el producto

4.9

Promedio entre 634 opiniones

5 estrellas	588
4 estrellas	40
3 estrellas	1
2 estrellas	1
1 estrella	4

Todas | Positivas | Negativa

Muy malo

Anda muy bien, pero tiene fallas de seguridad que no comparto, si acercas un busca polo se pende la luz, me dijeron que era una falla, la cambie por otra nueva y sucede lo mismo vengan 2 falladas con el mismo problema. Hace 5 meses Denunciar

5 9

No lo recomendaría de nuevo

Compre la misma batidora hace menos de un año. Dice tener garantía. Consulte y nadie me contesto. Para comprar y mandar son rápidos. Para responder ese tipo de consultas y arreglar no ya que no tuve respuesta. Más de 1 año

17 3

Muy malo

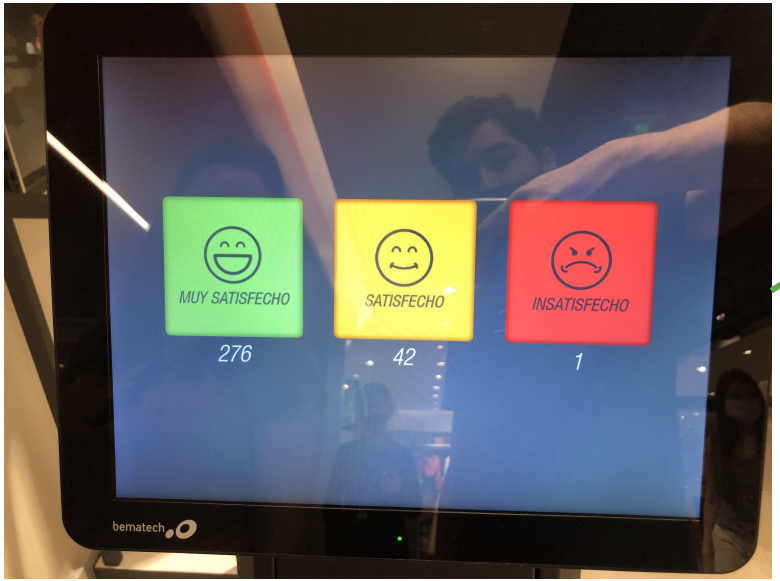
Producto defectuoso con faltantes de pieza. Y nadie se hace cargo porque tiene garantia de 2 dias!!!! cuando tarda una semana en llegar el producto!!. Más de 3 años

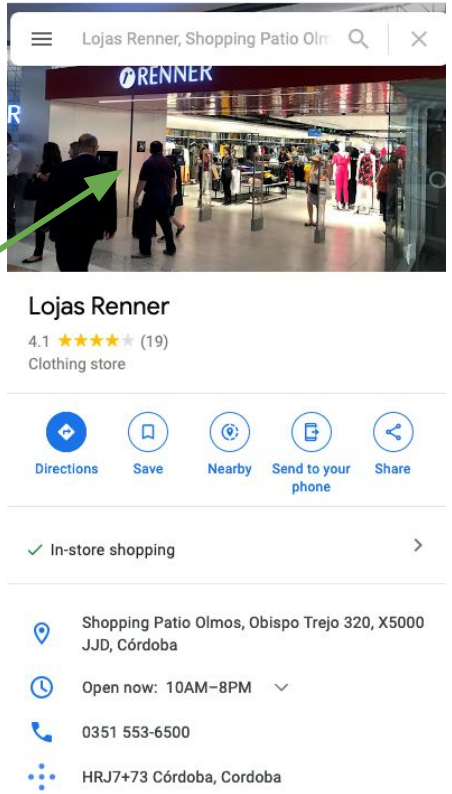
28 19

Insegura

Buen producto lástima que no controlan el aislamiento. Cuando la tocas da un poco la corriente. Confirmen con un busca polo. Más de 1 año

Sentiment Analysis - Ejemplos





Sentiment Analysis - Pero porque?

Dan Jurafsky

Why sentiment analysis?

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

Sentiment Analysis - Cómo medimos

Dan Jurafsky

Sentiment Analysis

- Simplest task:
  - Is the attitude of this text positive or negative?
- More complex:
  - Rank the attitude of this text from 1 to 5
- Advanced:
  - Detect the target, source, or complex attitude types

Sentiment Analysis - La solución para todo? :)

Recordando la primer presentación:

- El lenguaje humano es complejo y se encuentra evolucionando.
- Enseñar a una máquina que analice los matices gramaticales, slangs, errores de escrituras comunes. O enseñar a nuestro modelo que entienda como un determinado contexto afecta o no es complejo.