

Introduzione al Trattamento Automatico del Linguaggio - II

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore



DI COSA PARLEREMO

PARTE I

- Linguistica Computazionale (LC) e Trattamento Automatico del Linguaggio (TAL)
- Sfide del TAL
- Come processare il linguaggio
 - esempio di pipeline

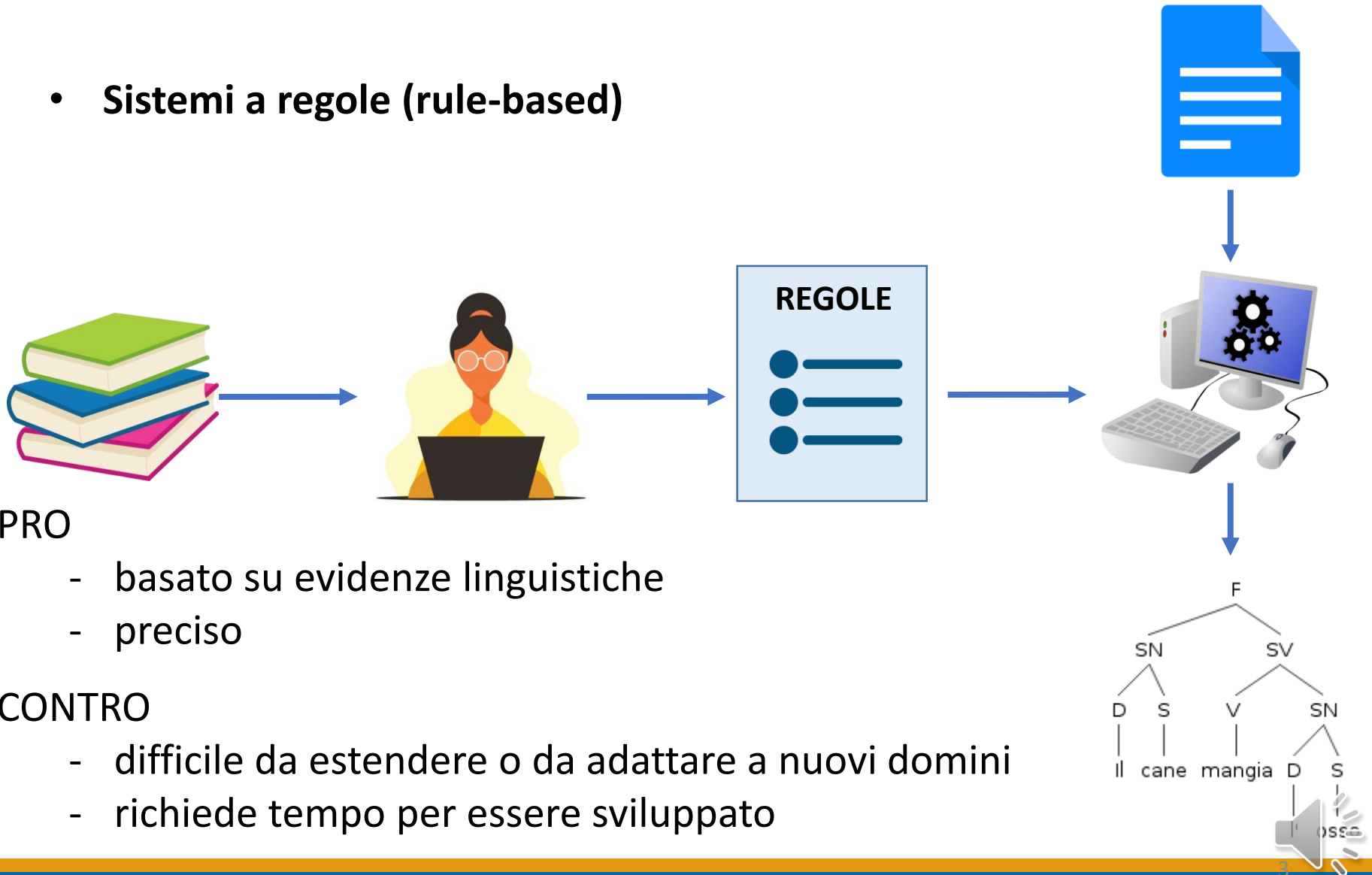
PARTE II

- Come si sviluppa un modulo di TAL
 - sistemi rule-based
 - sistemi machine learning:
 - approcci non supervisionati e supervisionati, annotazione, valutazione



COME SI SVILUPPA UN MODULO TAL

- Sistemi a regole (rule-based)



PRO

- basato su evidenze linguistiche
- preciso

CONTRO

- difficile da estendere o da adattare a nuovi domini
- richiede tempo per essere sviluppato

COME SI SVILUPPA UN MODULO TAL

- **Sistemi a regole (rule-based)**
- Esempio: Part-of-Speech tagging:
 - 1) assegnazione ad ogni parola di tutti i possibili PoS usando un dizionario

		NOUN
VERB	ART	VERB
«paghiamo	il	conto»

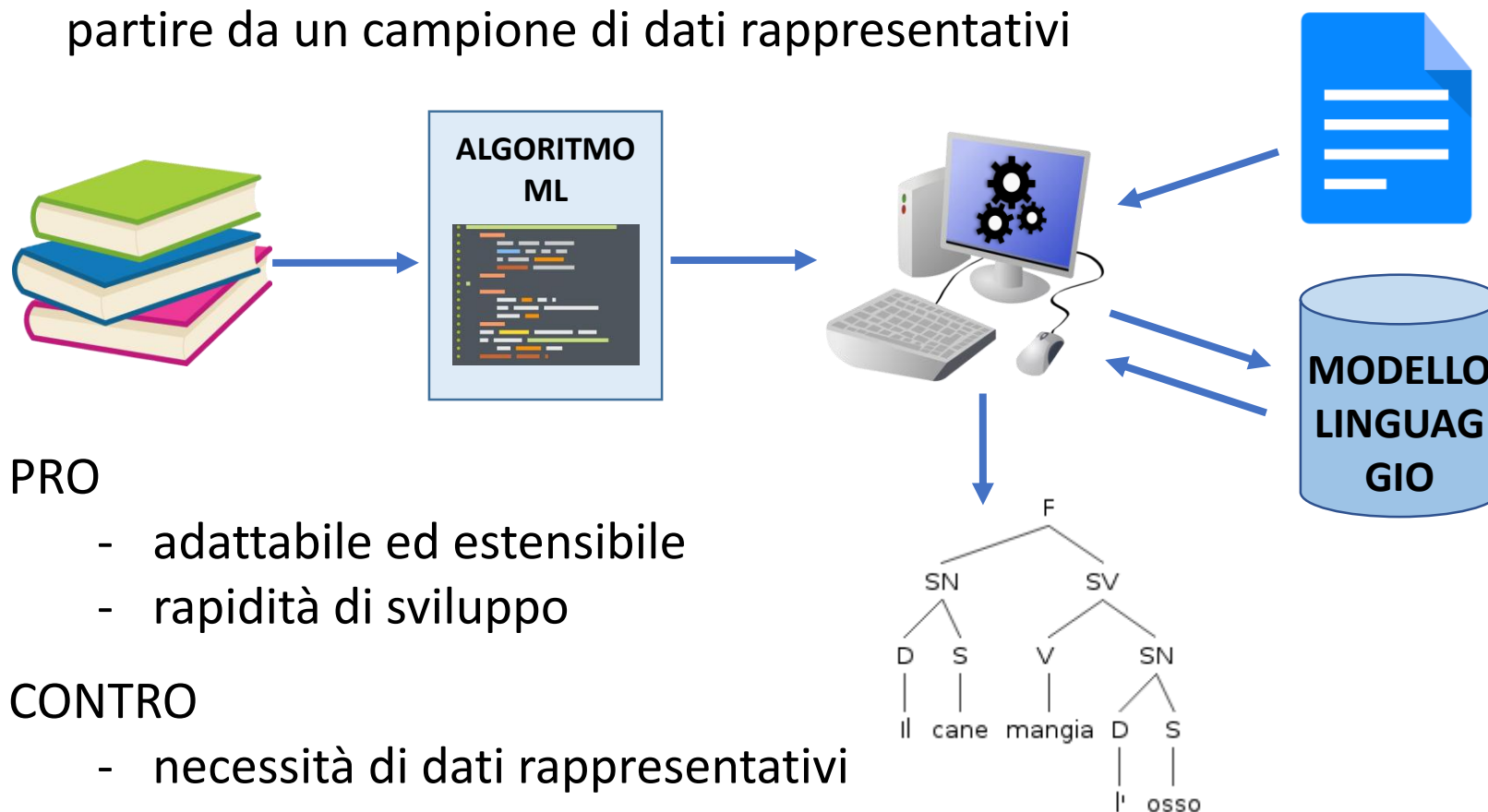
- 2) applicazione delle regole per rimuovere etichette ambigue
 - «rimuovere VERB se in alternativa con NOUN e preceduto da ART»

		NOUN
VERB	ART	VERB
«paghiamo	il	conto»



COME SI SVILUPPA UN MODULO TAL

- **Sistemi di apprendimento automatico – MACHINE LEARNING (ML)**
 - algoritmi che permettono al computer di imparare a svolgere un task a partire da un campione di dati rappresentativi



COME SI SVILUPPA UN MODULO TAL

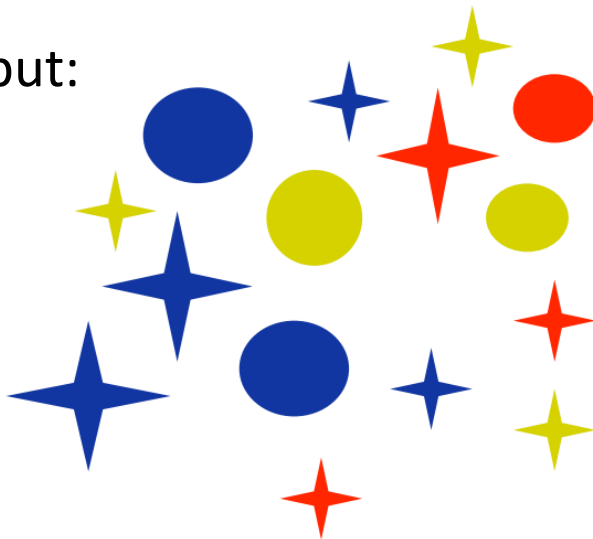
- **Sistemi di apprendimento automatico – MACHINE LEARNING (ML)**
- **3 tipi principali di algoritmi di ML**
 1. **NON SUPERVISIONATI:** non necessitano di un corpus annotato a mano per creare il modello
 2. **SUPERVISIONATI:** utilizzano un corpus annotato a mano per la creazione dei modelli
 3. **SEMI-SUPERVISIONATI:** combinano informazioni derivanti sia da corpora annotati che da dati non annotati



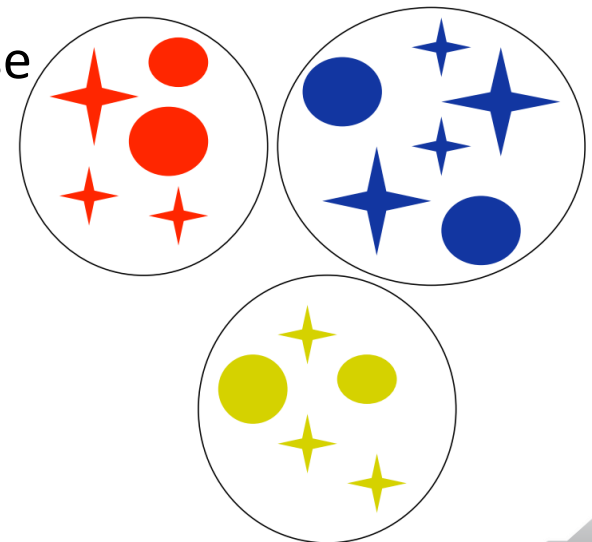
COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML NON SUPERVISIONATO, esempio
 - CLUSTERING: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:



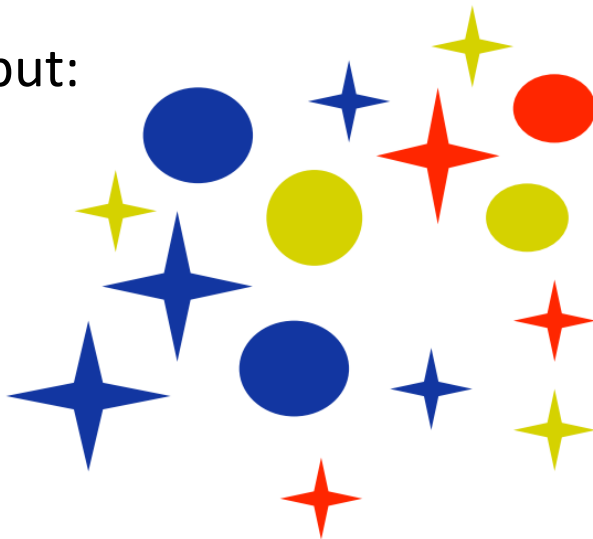
Output in base
al colore:



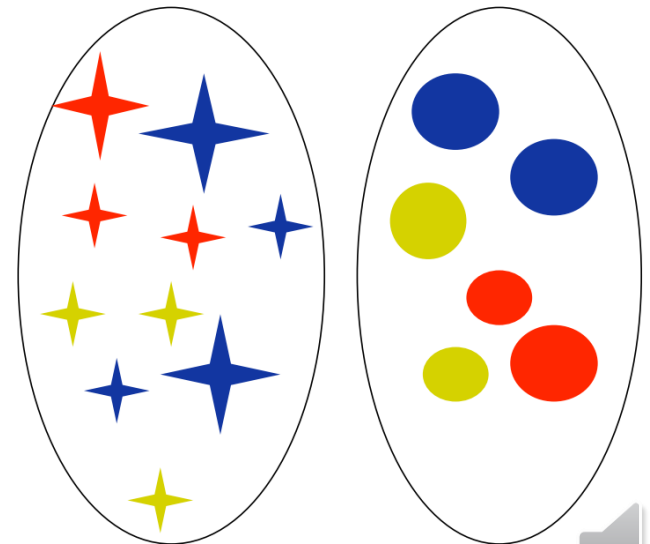
COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML NON SUPERVISIONATO, esempio
 - CLUSTERING: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:

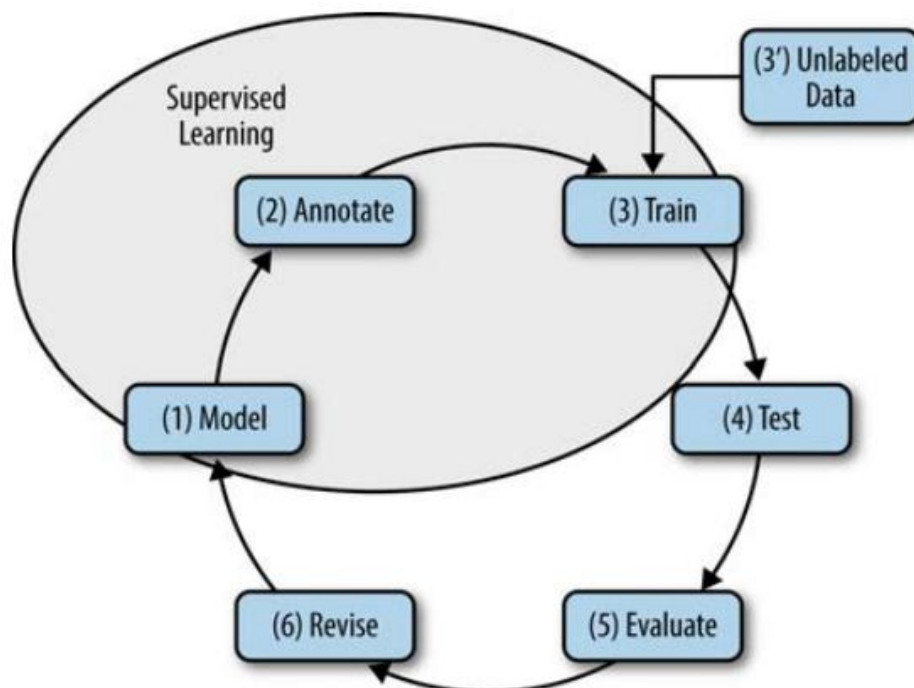


Output in
base alla
forma:



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO



Il ciclo MATTER

(Pustejovsky and Stubbs (2012) "Natural Language Annotation for Machine Learning". O'Reilly Media.)



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- Il ciclo MATTER:
 - **Model**: descrizione teorica di un fenomeno linguistico
 - **Annotate**: annotazione del corpus con uno schema di annotazione basato sul modello
 - **Train**: addestramento di un algoritmo di ML sul corpus annotato
 - **Test**: test del sistema addestrato su un nuovo campione di dati
 - **Evaluate**: valutazione delle performance del sistema
 - **Revise**: revisione del modello e dello schema di annotazione



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- ANNOTAZIONE
 - aggiunta di informazioni linguistiche al testo tramite etichette (*tag*)
 - copre ogni aspetto dell'analisi linguistica
 - rende esplicita e analizzabile dal computer la struttura linguistica implicita nel testo
- SCHEMA DI ANNOTAZIONE
 - repertorio di categorie linguistiche per l'annotazione: lista di tag e attributi
- LINEE GUIDA DI ANNOTAZIONE
 - documento in cui viene spiegato il *modo* in cui l'annotazione è proiettata sul testo



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO

- Esempio: ***Sentiment Polarity Classification***

subj	Subjectivity: possible values are 0 and 1. A subjective tweet will have subj = 1; an objective tweet subj = 0.
opos	Positive <i>overall</i> polarity: possible values are 0 and 1. A tweet exhibiting positive polarity will have opos = 1; a tweet without positive polarity will have opos = 0.
oneg	Negative <i>overall</i> polarity: possible values are 0 and 1. A tweet exhibiting negative polarity will have neg = 1; a tweet without negative polarity will have neg = 0.
iro	Irony: possible values are 0 and 1. A tweet with an ironic twist will have iro = 1, otherwise iro = 0.
lpos	Positive <i>literal</i> polarity: possible values are 0 and 1. A tweet exhibiting positive <i>literal</i> polarity will have pos = 1; tweet without positive <i>literal</i> polarity will have pos = 0.
lneg	Negative <i>literal</i> polarity: possible values are 0 and 1. A tweet exhibiting negative <i>literal</i> polarity will have neg = 1; tweet without negative <i>literal</i> polarity will have neg = 0.



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
 - Esempio: *Sentiment Polarity Classification*

subj	opos	oneg	iro	lpos	lneg	description and explanatory tweet in Italian
0	0	0	0	0	0	objective <i>l'articolo di Roberto Ciccarelli dal manifesto di oggi http://fb.me/1BQVy5Wak</i>
1	0	0	0	0	0	subjective with neutral polarity and no irony <i>Primo passaggio alla #strabrollo ma secondo me non era un iscritto</i>
1	1	0	0	1	0	subjective with positive polarity and no irony <i>splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura http://t.co/GWoZqbxAuS</i>
1	0	1	0	0	1	subjective with negative polarity and no irony <i>Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont... http://t.co/3CazKS7Y</i>
1	1	1	0	1	1	subjective with both positive and negative polarity (mixed polarity) and no irony <i>Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme" http://t.co/kIKnbFY7</i>
1	1	0	1	1	0	subjective with positive polarity, and an ironic twist <i>Questo governo Monti dei paschi di Siena sta cominciando a carburare; speriamo bene...</i>



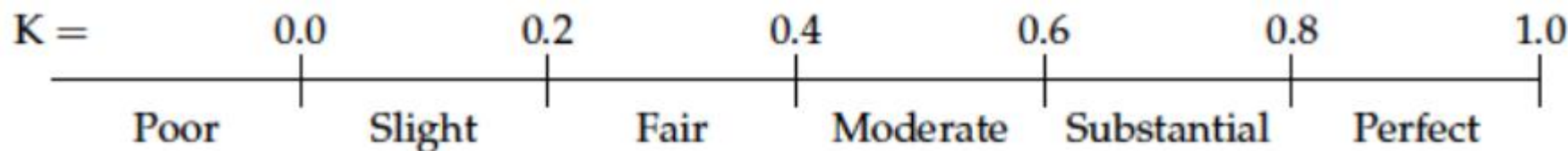
COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- Dati necessari:
 - di training (*training set*): dati annotati per l'addestramento del modello
 - di test (*test set*): dati NON annotati, diversi da quelli di training, su cui applicare il modello addestrato
 - di valutazione (*gold standard*): dati del test annotati su cui valutare le performance del modello addestrato



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **Inter-Annotator Agreement (IAA)** = accordo tra almeno 2 annotatori sullo stesso testo
 - consistenza dell'annotazione
 - plausibilità cognitiva del modello
 - un ampio accordo tra gli annotatori è considerato garanzia della validità di tale schema e dei dati annotati
 - K di Cohen (annotatori = 2) o di Fleiss (annotatori > 2)



COME SI SVILUPPA UN MODULO TAL

- **Sistemi di apprendimento automatico – MACHINE LEARNING (ML)**
- **ML SUPERVISIONATO**

TRAIN:

- Selezionare una serie di documenti da usare per addestrare l'algoritmo
- Annotare ciascun token nei documenti
- Identificare le feature appropriate
- Creare un classificatore che preveda le etichette

TEST:

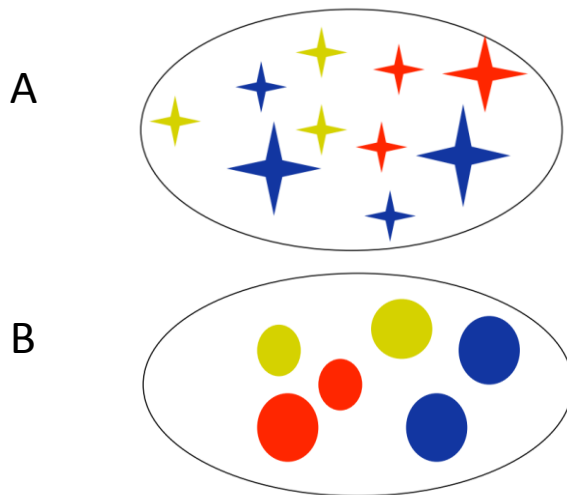
- Usare una serie di documenti non usati nel training
- Lanciare il classificatore per etichettare ciascun token
- Avere in output i documenti annotati
- Valutare sugli stessi documenti manualmente annotati



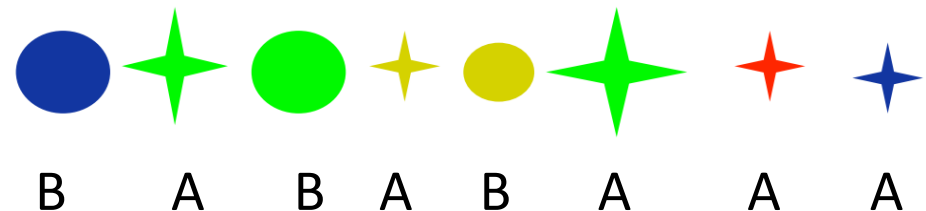
COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO, esempio
 - CLASSIFICAZIONE: dato un insieme di classi predefinite determinare a quale classe appartiene una certa entità

Input (training):

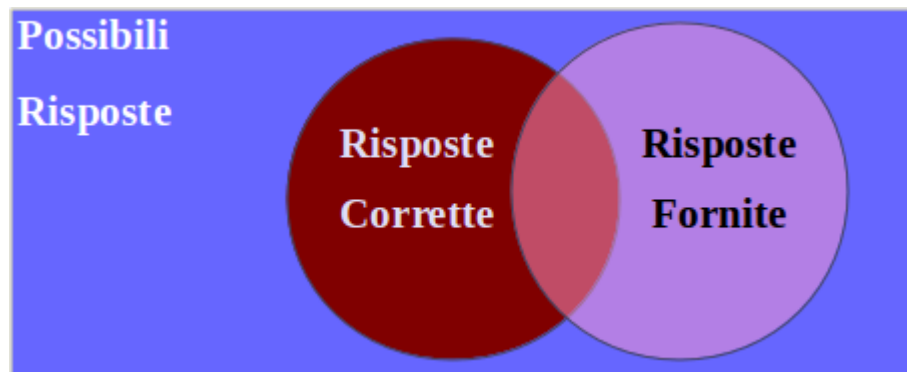


Classificazione di nuovi dati (test):



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - confronto dell'output del modello sui dati di test con il gold standard



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- VALUTAZIONE: matrice di confusione

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** matrice di confusione

	Positive	Negative	Total
Positive	72	271	343
Neutral	41	252	293
Negative	15	247	262
Total	128	770	898

Table 20. Confusion Matrix for Sentiment Analysis.

Figura tratta da «Social media sentiment analysis and topic detection for Singapore English»



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: ACCURACY

$$\text{ACCURACY} = \frac{\text{\#risposte corrette}}{\text{\#risposte fornite}}$$

Esempio:


- 150 frasi annotate nel test
- 120 frasi annotate con sentiment corretto
- accuracy = $120/150 = 0,8$ (80%)



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **PRECISION**, misura il rapporto tra le entità correttamente riconosciute dal sistema ed il totale delle entità riconosciute

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative




$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **RECALL**, misura il rapporto tra le entità correttamente riconosciute dal sistema ed il totale delle entità corrette

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative


$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - uso di metriche standard: **F-MEASURE**, media armonica tra precision e recall

$$\text{F-MEASURE} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE:** analisi quantitativa delle prestazioni del modello
 - Esempio:

		ACTUAL (gold standard)	
		Positive	Negative
PREDICTED (test set)	Positive	70 (TP)	15 (FP)
	Negative	30 (FN)	45 (TN)

- Precision: $70 / (70+15) = 70 / 85 = 0,82$

- Recall: $70 / (70+30) = 70 / 100 = 0,70$

- F-measure: $2 * 0,82 * 0,7 / (0,82 + 0,70) = 0,75$



COME SI SVILUPPA UN MODULO TAL

- Sistemi di apprendimento automatico – MACHINE LEARNING (ML)
- ML SUPERVISIONATO
- **VALUTAZIONE**: analisi quantitativa delle prestazioni del modello

	Precision	Recall	F-score
Positive	0.210	0.563	0.306
Negative	0.943	0.321	0.479

Table 21. Results for Sentiment Analysis using 898 Target Phrases. (Bolded entries are the best results for each row.)

Figura tratta da «Social media sentiment analysis and topic detection for Singapore English»

