

THE LEGEND CHALLENGE

PROGETTO SOASEC A.A. 2025

Professore:

Ernesto Damiani

Studenti:

Stucchi Jacopo 47294A

Bourras Ikram 53725A

Abstract:

Analisi di approcci differenti di fine-tuning per specializzare un modello pre-allenato su un determinato argomento, in particolare la *gender equality*.

SOMMARIO

Introduzione	4
Strategie utilizzate.....	5
<i>Strategia per convertire il testo della direttiva.....</i>	<i>5</i>
<i>Strategia per convertire il testo delle leggende.....</i>	<i>5</i>
Valutazione	7
<i>Adattamento del Glue benchmark.....</i>	<i>7</i>
Task selezionati e adattati	7
Text Classification	7
Bias Detection	7
Sentiment Analysis	7
Reading Comprehension	7
Struttura dei dataset.....	8
Classi previste per ciascun task.....	8
<i>Adattamento del SuperGlue benchmark.....</i>	<i>8</i>
Task selezionati e adattati	8
COPA (Choice of Plausible Alternatives).....	8
RTE (Recognizing Textual Entailment, in ottica legale)	9
MultiRC (Multiple-choice Reading Comprehension):	9
WiC (Word-in-Context):.....	9
Struttura dei dataset.....	9
Classi previste per ciascun task.....	9
COPA	9
RTE (LE)	9
MultiRC.....	9
WiC	10
<i>Valutazione dei modelli.....</i>	<i>10</i>
Analisi esplicativa.....	12
<i>Data-driven</i>	<i>12</i>
Glue evaluation	12
Text Classification	12
Sentiment Analysis	13
Bias Detection	13
Reading Comprehension	13
Miglior modello	14
SuperGlue evaluation	14
Choice of Plausible Alternatives (COPA):.....	14
Recognizing Textual Entailment (Legal Entailment):	15
Multi-Sentence Reading Comprehension:	16
Words in Context:	17
Miglior modello	18
<i>Model-driven.....</i>	<i>19</i>
Lime	19

Glue	19
Text classification	19
SuperGlue	21
RTE (LE)	21
MultiRC	21
WiC	22
Conclusioni	24

INTRODUZIONE

Il presente progetto universitario si propone di valutare e confrontare diversi modelli di linguaggio nell'elaborazione di quesiti astratti (normativi) e casi pratici (aziendali), mantenendo sempre coerenza con la normativa vigente. L'obiettivo è identificare quale modello risulti più efficace nel fornire risposte accurate e interpretabili.

A tal fine, sono stati realizzati due notebook Jupyter in Python per la valutazione dei task dei benchmark GLUE e SuperGLUE, con l'obiettivo di analizzare le performance dei modelli sia in termini di accuratezza che di interpretabilità. Tutti i materiali, inclusi i notebook, sono disponibili in versione open source su GitHub: https://github.com/JacoStu/SOASEC_project/tree/main.

La normativa di riferimento per il progetto è la Direttiva 2022/2381, disponibile sul sito ufficiale: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2381>.

STRATEGIE UTILIZZATE

STRATEGIA PER CONVERTIRE IL TESTO DELLA DIRETTIVA

Le direttive europee seguono una struttura ricorrente:

- **Preambolo:** spiega il contesto politico, giuridico e le motivazioni.
- **Corpo:** contiene gli obblighi normativi vincolanti, cioè gli articoli.
- **Conclusione:** stabilisce tempistiche e misure future.

Per analizzare la Direttiva 2022/2381, abbiamo suddiviso gli articoli in **sezioni logiche** e creato una **tabella di mapping** tra i punti del preambolo e gli articoli correlati, per mantenere coerenza e tracciabilità.

Sezione	Punti del Preambolo	Articoli Correlati
1. Introduzione / Contesto Politico	1, 5, 11	Art. 1, Art. 2
2. Riferimenti Giuridici / Normativi	2, 3, 4, 13, 14, 30, 38, 55	Art. 3,
3. Obiettivi della Direttiva	7, 9, 10, 16, 17, 21, 22, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 46, 47, 48, 49, 50	Art. 5, Art. 6, Art. 13
4. Destinatari e Soggetti Coinvolti	3, 8, 20	Art. 2, Art. 8,
5. Giustificazione della Necessità dell'Intervento Europeo	6, 8, 15, 18, 19, 23, 45, 52, 53	Art. 4, Art. 7, Art. 12
6. Tempistiche e Azioni Future	12, 54	Art. 11, Art. 14

Successivamente, sulla base di questa suddivisione, è stato definito un set di **ruoli e messaggi** per l'interazione con il modello:

- **System:** stabilisce il ruolo dell'assistente (esperto legale della direttiva) e le regole di comportamento nella risposta (verifica di applicabilità, sintesi, giustificazione normativa, suggerimenti operativi, gestione dei casi non previsti).
- **User:** contiene la domanda o il caso di interesse, formulato dal punto di vista dell'utente finale (ad esempio personale di selezione o azienda soggetta alla direttiva).
- **Assistant:** fornisce la risposta strutturata, rispettando le istruzioni definite nel ruolo system.

In questo modo si è costruito un flusso coerente e controllato, che permette di ottenere risposte affidabili, fondate sul testo normativo e orientate a fornire indicazioni operative concrete.

STRATEGIA PER CONVERTIRE IL TESTO DELLE LEGGENDE

Per arricchire il dataset di addestramento, è stata adottata una strategia narrativa basata sulla creazione di **cinque storie generate tramite un modello linguistico**.

Le storie avevano come protagonisti l'azienda e i soggetti che vi lavorano, e sono state progettate per coprire **scenari differenziati**:

- **Conformità:** casi in cui la direttiva (UE) 2022/2381 veniva rispettata correttamente, mostrando le pratiche positive da adottare.

- **Non conformità:** casi in cui la direttiva non veniva rispettata, includendo conseguenze e sanzioni previste, per insegnare al modello a riconoscere anche gli scenari negativi.

L'obiettivo è quello di trasformare il contenuto narrativo in **esempi supervisionati**, bilanciando casi positivi e negativi, così da permettere al fine-tuning di apprendere non solo le regole della direttiva, ma anche le implicazioni pratiche legate alla sua applicazione o violazione.

Anche nel caso delle leggende è stato definito un set di **ruoli e messaggi** per l'interazione con il modello:

- **System:** definisce il contesto: il modello deve comportarsi come un assistente legale esperto della Direttiva 2022/2381.
- **User:** contiene la domanda, legata a una **storia aziendale**, che serve da scenario realistico.
- **Assistant:** risponde in maniera aderente alla direttiva, ma usando lo scenario come spunto per spiegare la regola generale.

VALUTAZIONE

ADATTAMENTO DEL GLUE BENCHMARK

Per valutare i modelli allenati tramite fine-tuning sulla Direttiva Europea 2022/2381 e, più in generale, su temi legati alla parità di genere, è stato progettato un insieme di task ispirato al benchmark **GLUE (General Language Understanding Evaluation)**.

L'obiettivo è costruire benchmark ad hoc, specializzati nel verificare come i modelli allenati rispondessero in scenari rilevanti per il dominio della direttiva. I benchmark sono stati applicati a tre modelli distinti:

1. il **modello “no-ff”**, generale pre-allenato;
2. il **modello “directive”**, affinato secondo la strategia illustrata in precedenza;
3. il **modello “legends”**, affinato con racconti basati sulla direttiva, sempre seguendo la strategia descritta precedentemente.

Delle **nove** task originarie di GLUE, ne sono state selezionate **quattro**, adattandole al contesto della direttiva, in modo da ottenere una valutazione più mirata e completa.

TASK SELEZIONATI E ADATTATI

Il framework di valutazione mantiene quindi un legame concettuale con GLUE, ma introduce task personalizzati per misurare in modo specifico le competenze dei modelli nel dominio della Direttiva 2022/2381 e della parità di genere, garantendo un bilanciamento tra **abilità linguistiche generali** e **sensibilità tematica**.

TEXT CLASSIFICATION

Verifica la capacità del modello di riconoscere la posizione espressa in una frase rispetto alla direttiva (*Supportive, Neutral, Against*). Questo task combina aspetti di **MNLI** e **QNLI** di GLUE, legati alla classificazione delle relazioni tra testo e ipotesi.

BIAS DETECTION

Misura la sensibilità del modello verso frasi con contenuti distorti o imparziali riguardo alla parità di genere. È un adattamento del task di **CoLA** e, in parte, del **diagnostic set** di GLUE, orientato a rilevare usi linguistici inappropriati.

SENTIMENT ANALYSIS

Valuta come il modello interpreta il sentimento espresso verso la parità di genere, classificando frasi come *Positive* o *Negative*. Si ispira al task **SST-2 (Stanford Sentiment Treebank)** di GLUE, centrato sulla polarità delle opinioni.

READING COMPREHENSION

Testa la capacità del modello di rispondere correttamente a domande basate su paragrafi estratti dal testo della direttiva. È un adattamento di **RTE** e **QNLI**, che in GLUE misurano la comprensione testuale e la capacità di stabilire se una risposta è contenuta nel contesto.

STRUTTURA DEI DATASET

Ogni dataset è organizzato con la seguente struttura:

- **Sentence:** la frase di input
- **Gold_Label:** l'etichetta corretta (attesa)
- **Label_no_ft:** output del modello base
- **Label_directive:** output del modello allenato sulla direttiva
- **Label_legends:** output del modello allenato con racconti ispirati alla direttiva

CLASSI PREVISTE PER CIASCUN TASK

Ogni *Gold_Label* e ogni *Label* prodotte dai modelli sono organizzate secondo la seguente struttura:

- **Text Classification:** *Supportive, Neutral, Against*
- **Bias Detection:** *Biased, Not biased*
- **Sentiment Analysis:** *Positive, Negative*
- **Reading Comprehension:** *Yes, No*

ADATTAMENTO DEL SUPERGLUE BENCHMARK

Dopo i risultati estremamente positivi ottenuti con l'adattamento del GLUE benchmark che hanno dimostrato come tutti e tre i modelli:

1. il **modello "no-ft"**, generale pre-allenato;
2. il **modello "directive"**, affinato secondo la strategia illustrata in precedenza;
3. il **modello "legends"**, affinato con racconti basati sulla direttiva, sempre seguendo la strategia descritta precedentemente.

Avendo assimilato correttamente le questioni etiche e legali collegate alla tematica in esame, si è deciso di procedere anche con l'adattamento del **SuperGLUE**.

TASK SELEZIONATI E ADATTATI

L'adattamento di questi task al dominio normativo della parità di genere permette di valutare in modo più approfondito non solo la comprensione del testo, ma anche la capacità dei modelli di **inferire, ragionare e distinguere significati in contesti complessi**, aspetti fondamentali per l'applicazione dell'IA al diritto e alle policy europee.

COPA (CHOICE OF PLAUSIBLE ALTERNATIVES)

Valuta la capacità di un modello di **identificare la causa o l'effetto più plausibile in un determinato scenario**. È particolarmente utile per analizzare situazioni concrete legate all'applicazione della direttiva, verificando se il modello riesce a inferire correttamente le conseguenze di misure per la parità di genere.

RTE (RECOGNIZING TEXTUAL ENTAILMENT, IN OTTICA LEGALE)

Misura la capacità di stabilire se un testo **implica, contraddice o è indipendente rispetto a un'ipotesi**. L'adattamento a *legal entailment (LE)* consente di testare la comprensione normativa, chiedendo al modello di riconoscere se un articolo della direttiva implica o meno una determinata interpretazione giuridica.

MULTIRC (MULTIPLE-CHOICE READING COMPREHENSION):

Richiede di rispondere a domande basate su un testo in cui ciascuna risposta può essere vera o falsa indipendentemente dalle altre. Questo task si adatta bene alla complessità della direttiva, in cui le disposizioni normative possono contenere più elementi di risposta validi.

WIC (WORD-IN-CONTEXT):

Valuta se una stessa parola mantiene lo stesso significato in due contesti differenti. L'adattamento al nostro dominio consente di verificare come i modelli interpretano concetti chiave come "quota", "parità" o "rappresentanza" in articoli diversi, riducendo ambiguità semantiche.

STRUTTURA DEI DATASET

Poiché il **SuperGLUE** non è un benchmark tradizionale ma un insieme eterogeneo di task, ogni compito richiede una struttura di dataset specifica. Non è quindi possibile uniformare i dati in un unico schema, ma per ciascun task è stata definita una rappresentazione coerente con le sue caratteristiche.

CLASSI PREVISTE PER CIASCUN TASK

COPA

- **Premise**: frase di contesto
- **Choice_1 / Choice_2**: due alternative plausibili (causa o effetto)
- **Label**: scelta corretta [1, 2]
- **Label_no_ft, Label_directive, Label_legends**: output dei tre modelli

RTE (LE)

- **Premise**: testo di partenza (articolo normativo o estratto)
- **Hypothesis**: ipotesi da verificare
- **Label**: ["entailment", "not_entailment"]
- **Label_no_ft, Label_directive, Label_legends**: output dei tre modelli

MULTIRC

- **Passage**: testo di riferimento
- **Question**: domanda sul testo
- **Answer_1, Answer_2**: risposte candidate
- **Label_1, Label_2**: valori corretti [0, 1]
- **Label_1_no_ft, Label_1_directive, Label_1_legends**: versioni dei tre modelli per la risposta 1

- **Label_2_no_ft, Label_2_directive, Label_2_legends:** versioni dei tre modelli per la risposta 2

WIC

- **Sentence_1, Sentence_2:** frasi contenenti la stessa parola
- **Word:** parola da disambiguare
- **Label:** [0, 1] (0 = diverso significato, 1 = stesso significato)
- **Label_no_ft, Label_directive, Label_legends:** output dei tre modelli

VALUTAZIONE DEI MODELLI

Il primo passo è stato l'etichettatura dei dataset di valutazione.

Per i task del **GLUE** ogni frase è stata associata a una **Gold_Label**, che rappresenta il target per il calcolo delle metriche (accuracy, F1-score ecc.). Per i task del **SuperGLUE** è stato seguito lo stesso procedimento, utilizzando la colonna **Label** come corrispondente della Gold_Label.

Una volta preparati i dataset, ogni riga è stata sottoposta ai tre modelli in analisi (modello base non affinato, modello affinato sulla direttiva, modello affinato sulle storie "legends"). A ciascun modello è stato chiesto di restituire una label tra quelle previste dal task; le label predette sono state salvate in file CSV e messe a confronto con le Gold_Label per la valutazione quantitativa.

Nello specifico, per ogni riga del CSV sono presenti:

- la frase di input (Sentence / Premise / Passage / ...),
- la Gold_Label,
- la label restituita dal modello non affinato (**Label_no_ft**),
- la label restituita dal modello affinato sulla direttiva (**Label_directive**),
- la label restituita dal modello affinato sulle storie (**Label_legends**).

SYSTEM	Classify the stance about gender equality in listed companies boards directive. Answer with a label choosen among Supportive, Neutral, Against.	SYSTEM	Classify the stance about gender equality in listed companies boards directive. Answer with a label choosen among Supportive, Neutral, Against.
USER	EU Directive 2022/2381 is a crucial step towards ensuring equal opportunities in top management positions in European companies.	USER	EU Directive 2022/2381 is a crucial step towards ensuring equal opportunities in top management positions in European companies.
ASSISTANT	Supportive	ASSISTANT	Supportive

L'istantanea di **FineTuneDB** mostra come è stato strutturato il test: la istruzione di comportamento è fornita al ruolo **SYSTEM**, la frase da classificare è posta nel ruolo **USER** e la risposta effettiva del modello compare nel ruolo **ASSISTANT**. Questa procedura è stata ripetuta per tutte le righe e tutti i modelli, ottenendo così i CSV di valutazione finali.

Questi CSV alimentano due fasi successive:

1. **Valutazione quantitativa** — calcolo di metriche aggregate (accuracy, F1, MCC, EM, ecc.) e costruzione di matrici di confusione per identificare pattern di errore.
2. **Analisi esplicativa** — studio delle ragioni delle decisioni del modello con approcci data-driven (analisi degli errori, wordcloud, n-gram, correlazioni con lunghezza del testo) e model-driven (LIME): per questo ultimo approccio abbiamo simulato una funzione

predict_proba basata sulle label predette e usato LIME per ottenere spiegazioni locali, concentrandoci in particolare sui casi errati.

Questo workflow (annotazione → esecuzione modelli → salvataggio CSV → analisi quantitativa + spiegazioni) permette non solo di misurare l'efficacia dei modelli, ma anche di motivare *perché* una certa predizione è stata scelta o perché si è verificato un errore, fornendo così evidenze utili per giustificare la scelta del modello più adatto al dominio della direttiva.

ANALISI ESPLICATIVA

Nell'ambito dell'**Explainable AI** si distinguono due approcci principali: **model-driven** e **data-driven**. Il primo utilizza le informazioni interne al modello per comprenderne il funzionamento, mentre il secondo si concentra esclusivamente sulle relazioni input-output, costruendo spiegazioni a partire dai dati senza accedere alla struttura interna.

DATA-DRIVEN

GLUE EVALUATION

Per ogni task, è stata calcolata l'**accuratezza** e il **coefficiente di correlazione di Matthews (MCC)**, confrontando le predizioni dei modelli con le etichette di riferimento presenti nei dataset. Sono stati generati i **word cloud**, per il task di **Text Classification**, per rappresentare visivamente gli errori commessi dai modelli, fornendo un'ulteriore prospettiva sui token che hanno causato difficoltà.

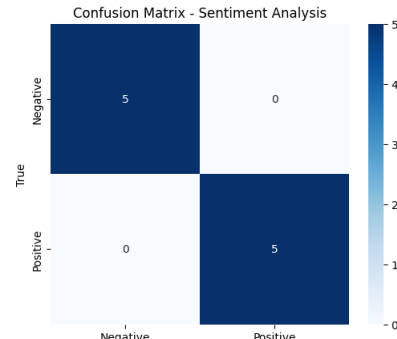
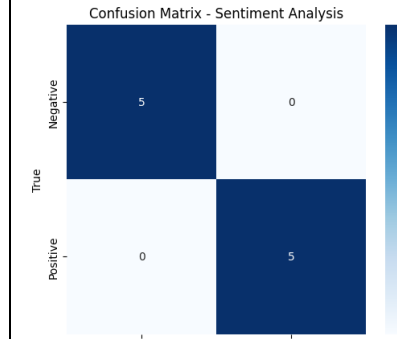
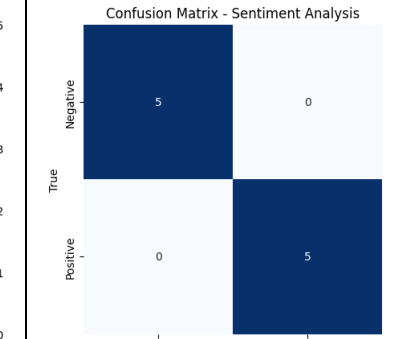
TEXT CLASSIFICATION

I risultati mostrano che tutti i modelli raggiungono performance elevate, ma evidenziano criticità specifiche. Le matrici di confusione confermano che le classi *Against* e *Supportive* vengono classificate correttamente, mentre la classe *Neutral* rimane la più problematica, con frequenti confusioni verso posizioni favorevoli o contrarie. L'analisi dei word cloud rivela che termini fortemente marcati, come **Gender**, **Balance** o **Directive**, tendono a orientare il modello verso etichette **polarizzate**, anche in contesti neutrali, mentre parole più ambigue come **may**, **companies**, **difficult**, generano incertezza. Nel complesso, i modelli risultano affidabili e spiegabili: gli errori non sono casuali ma riconducibili a pattern semantici specifici, offrendo così indicazioni concrete per futuri miglioramenti.

Model no fine-tuning	Model fine-tuning directive	Model fine-tuning legends
Accuracy 0.889	Accuracy 0.889	Accuracy 0.889
MCC 0.849	MCC 0.849	MCC 0.849

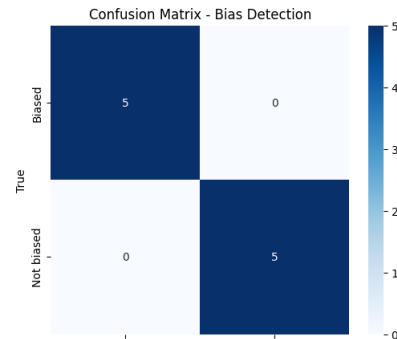
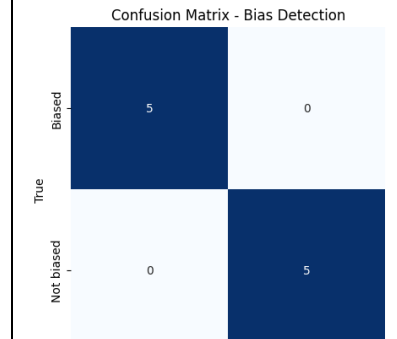
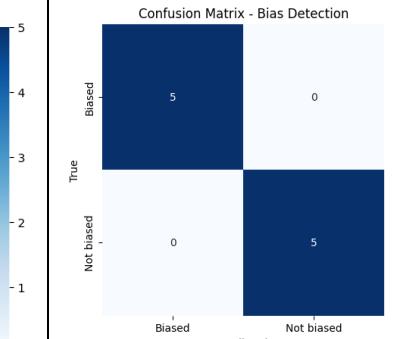
SENTIMENT ANALYSIS

Tutti i modelli hanno raggiunto prestazioni perfette. In questo caso non emergono problematiche di classificazione, confermando la piena affidabilità del modello per questo compito.

Model no fine-tuning	Model fine-tuning directive	Model fine-tuning legends
Accuracy 1.000	Accuracy 1.000	Accuracy 1.000
<div>Confusion Matrix - Sentiment Analysis</div> 	<div>Confusion Matrix - Sentiment Analysis</div> 	<div>Confusion Matrix - Sentiment Analysis</div> 

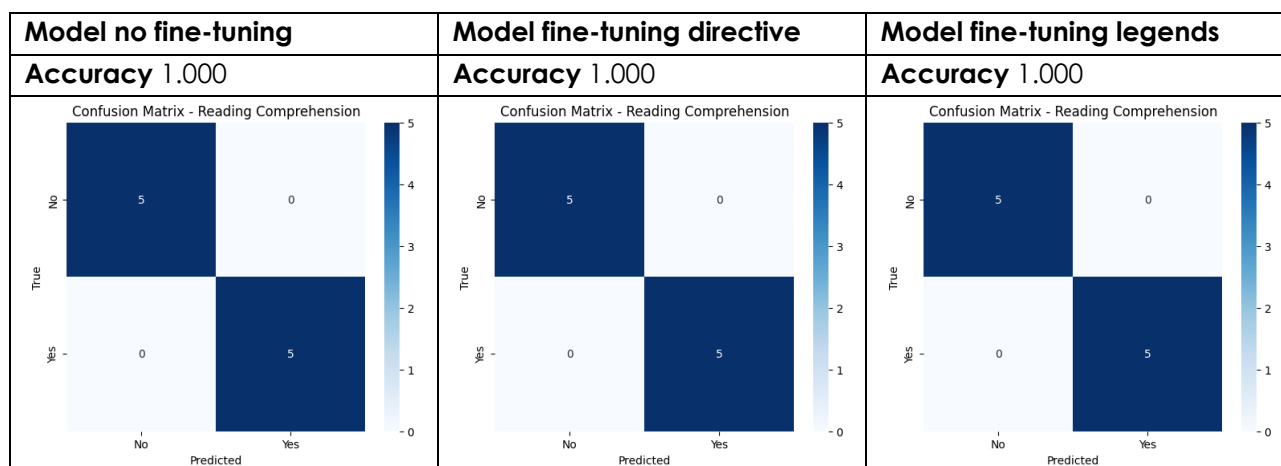
BIAS DETECTION

Tutti i modelli hanno raggiunto prestazioni perfette. In questo caso non emergono problematiche di classificazione, confermando la piena affidabilità del modello per questo compito.

Model no fine-tuning	Model fine-tuning directive	Model fine-tuning legends
Accuracy 1.000	Accuracy 1.000	Accuracy 1.000
<div>Confusion Matrix - Bias Detection</div> 	<div>Confusion Matrix - Bias Detection</div> 	<div>Confusion Matrix - Bias Detection</div> 

READING COMPREHENSION

Tutti i modelli hanno raggiunto prestazioni perfette. In questo caso non emergono problematiche di classificazione, confermando la piena affidabilità del modello per questo compito.



MIGLIOR MODELLO

La **macro-media** è stata calcolata considerando tutte le metriche disponibili per i vari task (accuracy, MCC e altre ove applicabili), in modo da fornire una valutazione unitaria e bilanciata. Dall'analisi emerge che **tutti e tre i modelli hanno raggiunto valori molto elevati e pressoché equivalenti**. Questo indica che le prestazioni complessive sono solide indipendentemente dal modello, ma i modelli con fine-tuning risultano comunque più **specializzati e contestualizzati** al dominio della direttiva, rendendoli preferibili per applicazioni pratiche reali, nonostante le metriche globali non mostrino differenze sostanziali.

	Macro-accuracy average	Macro average
Model no fine-tuning	0.972	0.977
Model fine-tuning directive	0.972	0.977
Model fine-tuning legends	0.972	0.977

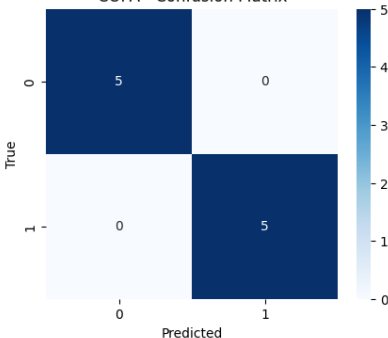
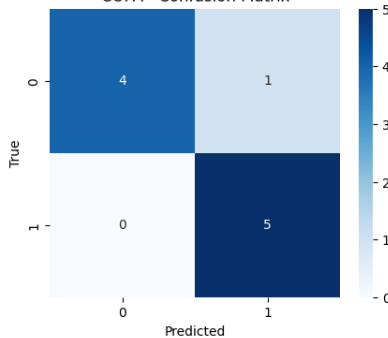
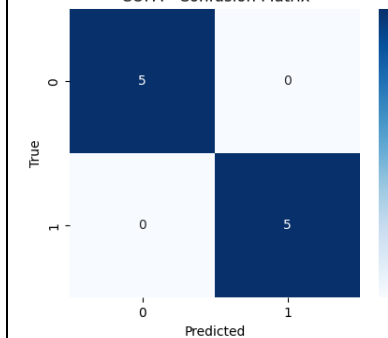
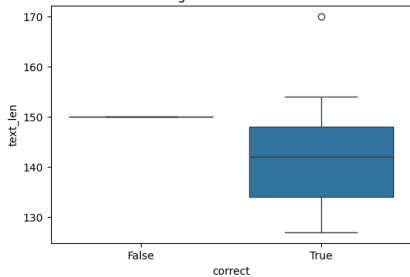

SUPERGLUE EVALUATION

Per ogni task, sono state calcolate le metriche di performance come l'**accuratezza**, **F1-score**, **MCC** ed **Exact Match (EM)**, confrontando le predizioni dei modelli con le etichette di riferimento.

CHOICE OF PLAUSIBLE ALTERNETIVES (COPA):

Tutti i modelli hanno raggiunto prestazioni perfette, senza particolari problematiche di classificazione, confermando la piena affidabilità sul benchmark SuperGLUE.

Unica eccezione è un errore riscontrato nel task COPA, dove il modello ha selezionato l'alternativa sbagliata. Dall'analisi tramite wordcloud, si osserva che i termini **“regulator”**, **“published”**, **“new”**, **“guidelines”** e **“directive”** hanno influito negativamente sulla predizione, inducendo il modello a confondere l'opzione corretta.

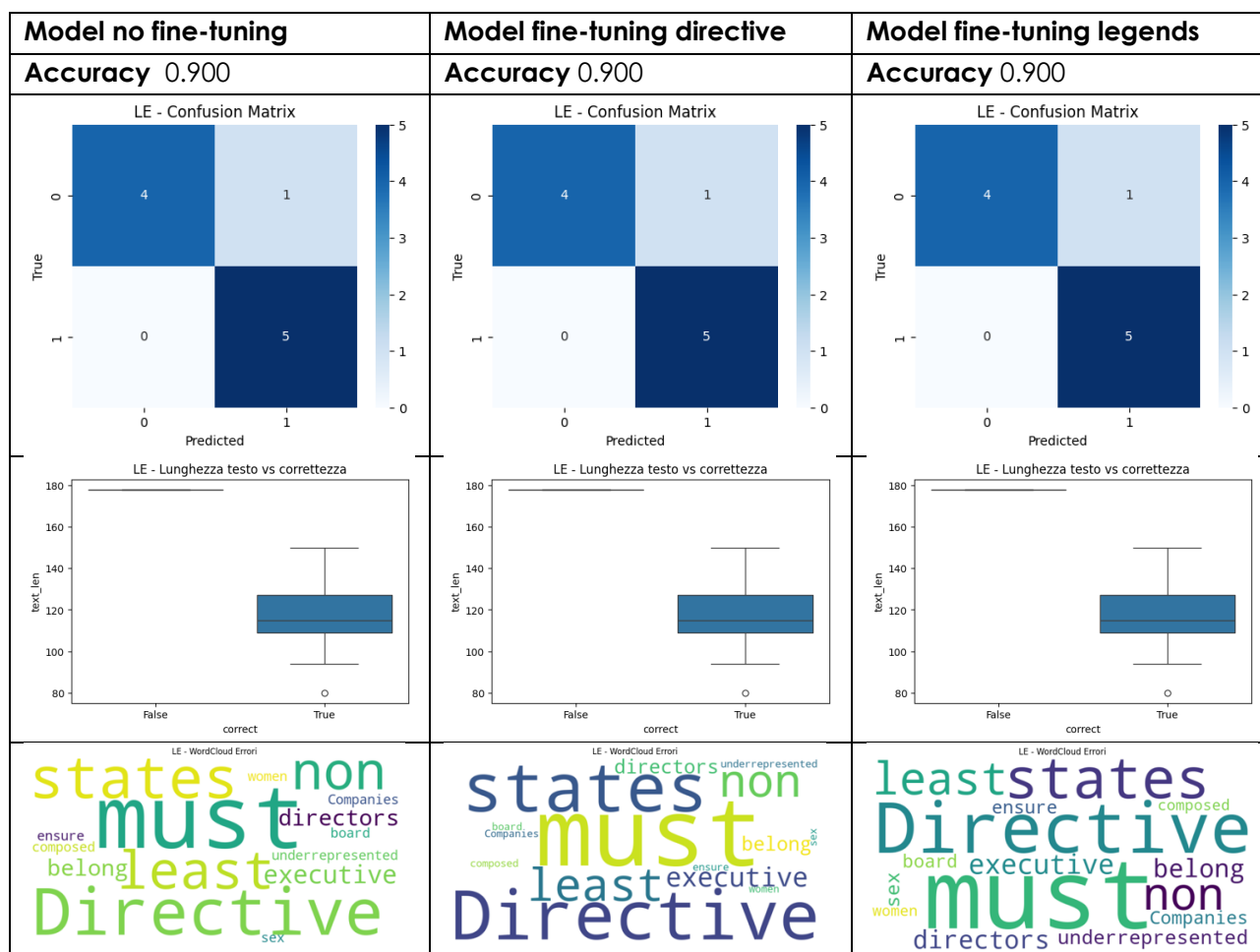
Model no fine-tuning	Model fine-tuning directive	Model fine-tuning legends
Accuracy 1.000	Accuracy 0.900	Accuracy 1.000
<p>COPA - Confusion Matrix</p> 	<p>COPA - Confusion Matrix</p> 	<p>COPA - Confusion Matrix</p> 
	<p>COPA - Lunghezza testo vs correttezza</p> 	
	<p>COPA - WordCloud Errori</p> 	

RECOGNIZING TEXTUAL ENTAILMENT (LEGAL ENTAILMENT):

Dalle matrici di confusione emerge che tutti e tre i modelli hanno commesso un errore di classificazione sullo stesso esempio del dataset RTE.

L'analisi tramite **boxplot** ha evidenziato come gli errori si concentrino su input particolarmente lunghi, con una distribuzione intorno ai 180 token, suggerendo che la complessità testuale possa aver influito negativamente sulla predizione.

Il **wordcloud** generato ha messo in evidenza i token che hanno maggiormente contribuito all'errore di classificazione ovvero **must, directive, states, least, non**. Questi termini, centrali per l'interpretazione semantica del testo, hanno probabilmente aumentato l'ambiguità del modello nel distinguere correttamente tra *entailment* e *not entailment*.




MULTI-SENTENCE READING COMPREHENSION:

Nel task è stato osservato un errore di classificazione nella seconda risposta alla domanda, dove il modello **“directive”** ha predetto **False** invece che **True**. L'analisi tramite wordcloud ha mostrato che i token **directive, balance, boards, promotes, gender, improve** hanno avuto un peso determinante nella predizione errata, aumentando l'ambiguità semantica tra la finalità generale della direttiva e la qualità decisionale.

Il **boxplot** della lunghezza dei testi ha evidenziato che gli errori si collocano tipicamente intorno a 150 token, mentre i testi correttamente classificati si distribuiscono tra i 100 e i 200 token, con mediana attorno a 140 e valori compresi nel range [130, 170]. Questo suggerisce che input più lunghi e complessi possano incrementare la probabilità di errore.

Dal confronto delle metriche si osserva che: Il **modello senza fine-tuning** e quello **sulle leggende** hanno raggiunto performance perfette invece, Il **modello sulla direttiva** ha mostrato prestazioni leggermente inferiori, indicando qualche difficoltà nel generalizzare correttamente le risposte, pur mantenendo comunque un livello molto elevato di accuratezza.

Model no fine-tuning	Model fine-tuning directive	Model fine-tuning legends
F1: 1.000	F1: 0.967	F1: 1.000
Em_score 1.000	Em_score 0.900	Em_score 1.000
		
		
		

WORDS IN CONTEXT:

Nel task sono stati osservati alcuni errori di classificazione legati all'ambiguità semantica di termini chiave come **respect**, **gender** e **suffered**. Nell'esempio con il termine **respect**, il modello ha confuso l'uso del concetto in un contesto di relazioni lavorative con quello legato ai principi di diversità, influenzato da token come **directive**, **gender**, **shareholders**. Analogamente, con **gender**, il modello non ha distinto tra l'uso normativo (quote di genere) e quello colloquiale (sesso biologico), generando un'interpretazione errata. Infine, con **suffered**, l'errore è stato indotto dall'associazione al contesto sanzionatorio, amplificata da token come **directive**, **employee**, **balance**.

Le metriche mostrano che il **modello senza fine-tuning** ha ottenuto risultati perfetti. I modelli **sulla direttiva** e **sulle leggende** hanno ottenuto prestazioni inferiori, con errori ripetuti sugli stessi input complessi.

L'analisi tramite **boxplot** ha evidenziato che gli errori si concentrano su testi più lunghi. Per il modello affinato sulla direttiva, i testi classificati correttamente avevano una lunghezza media tra 130 e 140 token, mentre gli errori si collocavano tra i 150 e i 180 token, con mediana sopra i 160. Per il modello affinato sulle leggende si osserva un pattern simile: i casi corretti hanno una mediana più bassa (± 135 token), mentre i falsi positivi si concentrano su input lunghi (160 -180 token).

Il task evidenzia come l'aumento della lunghezza dei testi e l'ambiguità lessicale possano ridurre la precisione dei modelli, soprattutto di quelli con fine-tuning, confermando la necessità di un'attenzione particolare alla gestione dei termini complessi.



MIGLIOR MODELLO

È stata calcolata una **macro-media** delle metriche per confrontare le **performance complessive** dei modelli. Il modello senza fine-tuning è il migliore, quello con fine-tuning su *legends* resta valido, mentre quello su *directive* riduce l'accuratezza complessiva

	Macro average
Model no fine-tuning	0.975
Model fine-tuning directive	0.833
Model fine-tuning legends	0.925

MODEL-DRIVEN

Per l'analisi **model-driven** è stata utilizzata la libreria **LIME** (*Local Interpretable Model-agnostic Explanations*).

Abbiamo scelto questo approccio perché i modelli considerati sono **black-box**, quindi non permettono un'interpretazione diretta del processo decisionale.

LIME

LIME genera **spiegazioni locali** delle predizioni, approssimando il comportamento del modello in prossimità dell'istanza analizzata. È stata simulata una funzione **predict_proba** che restituisce probabilità *one-hot* in base all'etichetta predetta. L'analisi è stata condotta **solo sugli errori di classificazione**, filtrando il dataset in modo da concentrarsi sui casi critici.

GLUE

Per l'analisi del **GLUE** abbiamo analizzato solo il **text-classification**, task in cui i modelli hanno commesso errori di classificazione.

TEXT CLASSIFICATION

I grafici mostrano le feature più rilevanti identificate dai tre modelli per ciascuna delle classi: **Supportive**, **Neutral** e **Against**.

- **Supportive**: i tre modelli hanno individuato pattern simili, con una forte enfasi su termini legati alla direttiva e al contesto aziendale. Le differenze principali emergono nell'assegnazione dei pesi ai token marginali.
- **Neutral**: qui si osservano variazioni più marcate tra i modelli. Il modello senza fine-tuning tende a distribuire pesi negativi a più token, mentre i modelli specializzati (*directive* e *legends*) mostrano una maggiore stabilità e coerenza semantica.
- **Against**: la divergenza tra modelli è più evidente. Il modello *no fine-tuning* assegna pesi positivi diffusi, mentre i modelli affinati tendono ad assegnare pesi più negativi ai token.

Errori di classificazione:

Caso 1: target label = *Neutral*

- modello **no-ft** -> *Against*

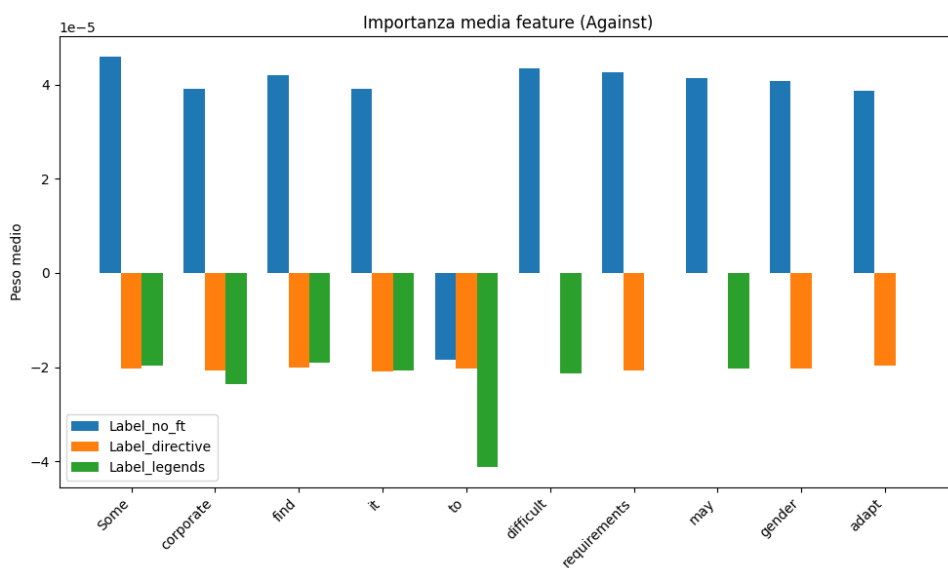
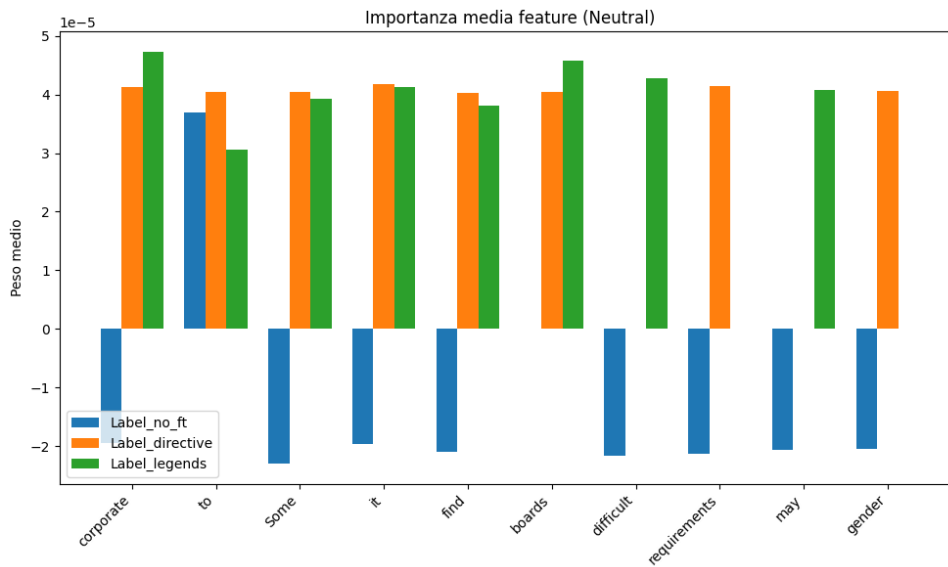
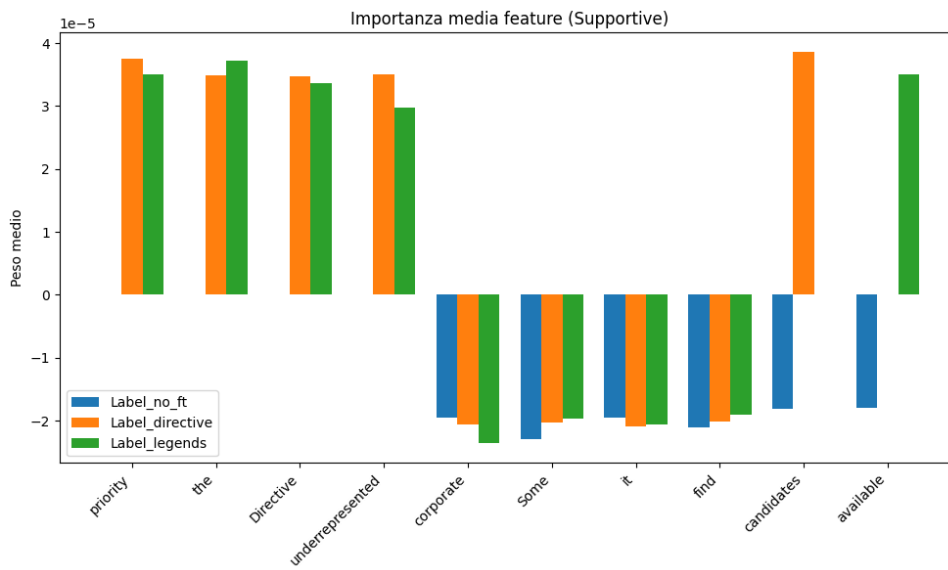
Il modello non affinato ha classificato come **Against**, quando avrebbe dovuto classificare come **Neutral**. Guardando i grafici della distribuzione dei *pesi dei token su ogni label*, si nota come il modello non affinato assegni pesi negativi ai token sia per la label **Supportive** che per la label **Neutral**, andando quindi nella direzione di **Against**.

Caso 2: target label = *Neutral*

- modello **directive** -> *Supportive*
- modello **legends** -> *Supportive*

In questo caso, i modelli affinati sulla direttiva e sulle storie, classificano come **Supportive** invece di **Neutral**. Guardando ai grafici si nota come i due modelli assegnino principalmente pesi positivi ai

token per le label **Supportive** che **Neutral**, mentre assegnano solo pesi negativi alla label **Against**, andando quindi a classificare quasi sempre come **Supportive** o **Neutral**.



SUPERGLUE

Per l'analisi del SuperGLUE abbiamo analizzato tutti i task escluso il COPA, task in cui i modelli hanno commesso errori di classificazione.

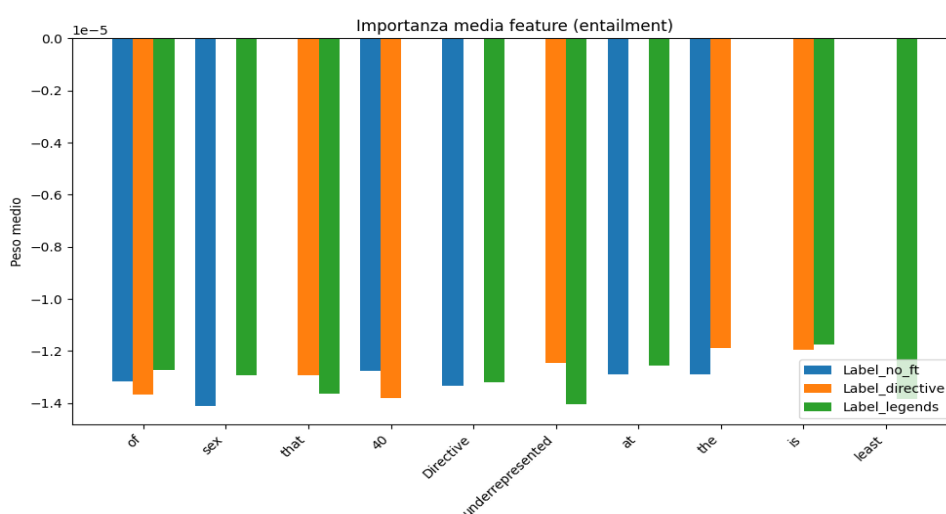
RTE (LE)

In questo caso, **tutti e tre i modelli hanno commesso lo stesso errore di classificazione**, non riuscendo a identificare correttamente la relazione di **entailment**. Il grafico mostra come ciascun modello ha pesato le parole più rilevanti del testo. Si osserva che tutti i pesi sono **negativi**, indicando che le features considerate hanno contribuito in modo sfavorevole alla classificazione. Il grafico evidenzia che i tre modelli hanno dato importanza simile alle stesse parole chiave, ma con sfumature diverse. Queste variazioni non sono state sufficienti a migliorare la classificazione: l'errore è rimasto comune a tutti i modelli, suggerendo che la difficoltà sia legata più al **contenuto semantico complesso della frase** che alla capacità del modello di cogliere singole parole.

Errore di classificazione:

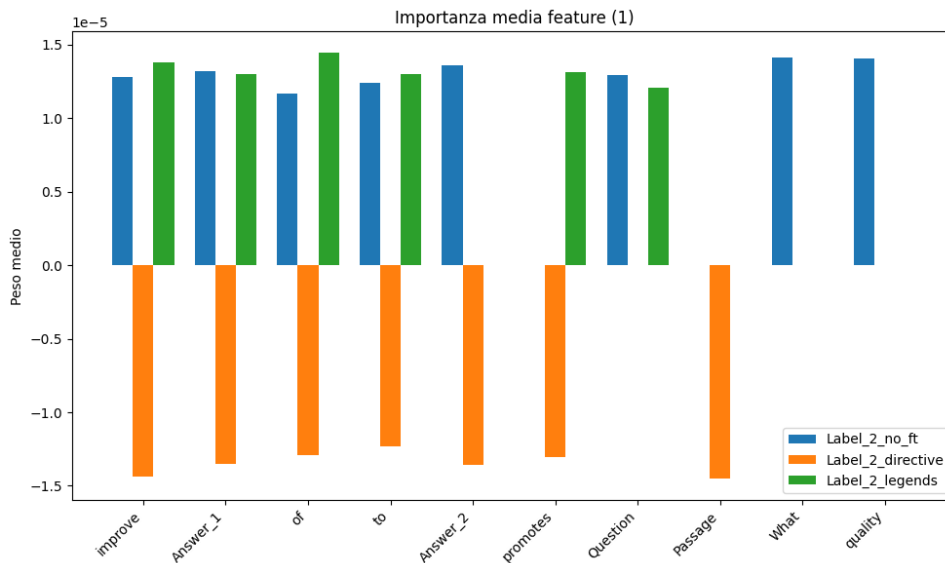
Target label = *entailment*

- tutti e tre i modelli -> *not-entailment*



MULTIRC

I modelli dovevano distinguere tra due label. I modelli hanno performato bene sulla *label_1*, mentre per la *label_2* il modello **directive** ha commesso errori. Dal grafico emerge infatti che questo modello interpreta alcune parole in modo **opposto** rispetto agli altri due, attribuendo loro un peso di segno inverso. Ciò suggerisce che il fine-tuning specifico sulla direttiva abbia modificato radicalmente l'associazione tra parole e predizioni, generando un'interpretazione non coerente con quella appresa dal modello base o dal modello affinato sulle *legends*. In questo contesto, la **direzione del peso (positivo vs negativo)** diventa un fattore chiave per comprendere come le parole influenzino le decisioni del modello.



WIC

L'obiettivo era verificare se una stessa parola avesse o meno lo stesso significato in due frasi. L'errore dei modelli non è stato causato direttamente dalla *word label* fornita, ma piuttosto da altre parole presenti nelle frasi, come **gender** o **directive**, fortemente legate al dominio delle direttive. Tali parole hanno esercitato un **peso negativo** nelle predizioni, spingendo i modelli verso una classificazione errata.

Dall'analisi si nota che tutti e tre i modelli sono stati influenzati in modo simile, ma con maggiore intensità nei modelli **directive** e **legends**, dove parole come **gender**, **company** e **directive** hanno assunto pesi negativi significativi. Questo ha inciso pesantemente sulle prestazioni, portando a una maggiore frequenza di errori rispetto al modello senza fine-tuning.

Caso 1: target label = 1

Word: *respect*

- modello **directive** -> 0
- modello **legends** -> 0

In questo caso, guardando al grafico, si nota come i token **gender**, **balance** e **directive** abbiano pesi negativi per la label 1 nei modelli affinati sulla direttiva e sulle storie, inducendo quindi i due modelli a classificare come 0, invece di 1.

Caso 2: target label = 0

Word: *gender*

- modello **directive** -> 1

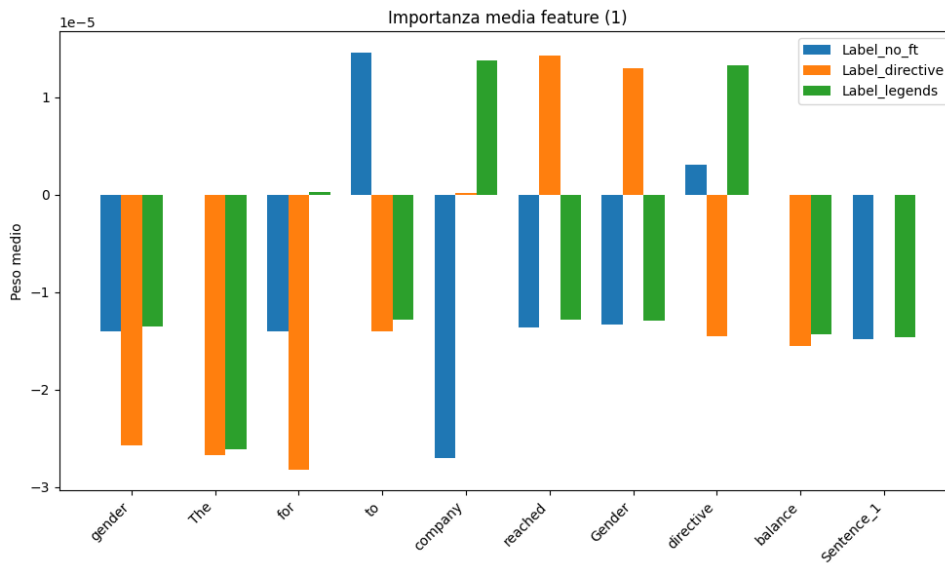
In questo caso, guardando al grafico, si nota come i token **reached** e **Gender** abbiano pesi positivi per la label 1 nel modello **directive**, inducendolo quindi a classificare come 1, invece di 0.

Caso 3: target label = 0

Word: *suffered*

- modello **legends** -> 1

In questo caso, guardando al grafico, si nota come i token **company** e **directive** abbiano pesi positivi per la label 1 nel modello legends, inducendolo quindi a classificare come 1, invece di 0.



CONCLUSIONI

L'utilizzo combinato dei benchmark **GLUE** e **SuperGLUE** insieme al metodo **LIME** ha permesso di confrontare i tre modelli non solo in termini di **accuratezza**, ma anche di **interpretabilità**, evidenziandone punti di forza e debolezze. L'uso di LIME ha fornito un livello aggiuntivo di **trasparenza**, rendendo possibile comprendere meglio le ragioni alla base delle decisioni dei modelli e valutare l'efficacia del **fine-tuning** basato sulla **direttiva** e sulle **storie** contestualizzate.

In generale, il modello base non allenato performa bene, ma quando si entra nel dettaglio della gender equality tende a mostrare limiti, in particolare nel riconoscere parole chiave come **risk**, **gender**, **equality** e **requirement**, che assumono un significato specifico nel contesto legale. I due modelli su cui è stato effettuato il fine-tuning performano meglio in questo scenario. Tra di essi, la differenza di performance non è elevata, ma il modello addestrato con le **legends** mostra risultati leggermente superiori. Questo è dovuto al fatto che i modelli apprendono più efficacemente dagli **esempi concreti** che dagli articoli, grazie alla maggiore diversità linguistica e alla ricchezza delle storie, che presentano scene differenti pur mantenendo lo stesso soggetto, garantendo così una maggiore ampiezza di contesto.