Classificazione automatica delle emozioni con BERT applicata a Stack Overflow
Emotion detection with BERT applied to Stack Overflow

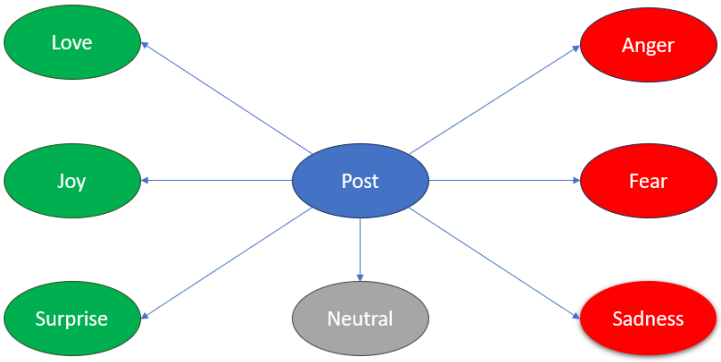
Relatore: Prof. Michele Tomaiuolo
Riassunto della tesi di laurea di: Matteo Gianvenuti, Matricola: 321490

Riassunto

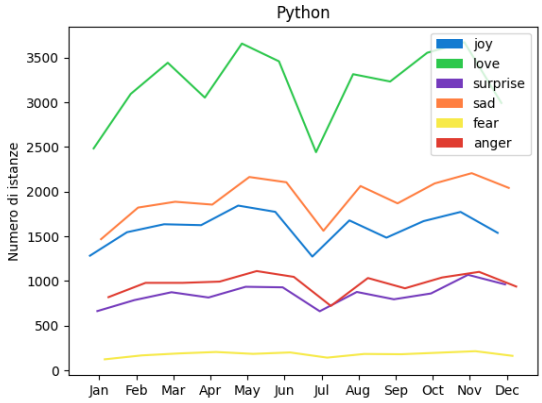
Negli ultimi anni l'analisi del sentimento è diventata sempre più importante specialmente nel campo dell'intelligenza artificiale e delle scienze sociali grazie al ruolo che ricopre nella società. L'opinione di altre persone, in particolare l'opinione pubblica, è sempre più influente nelle scelte di un individuo. Questo andamento riguarda anche i linguaggi di programmazione. È quindi interessante capire quale è il sentiment espresso verso i linguaggi di programmazione. L'obbiettivo finale è quello di osservare come variano le emozioni espresse verso i linguaggi di programmazione nel corso dell'anno 2022.
Per capire quale è il sentiment espresso verso i linguaggi di programmazione è necessario analizzare un'elevata quantità di pubblicazioni riguardanti quest'ultimi. Per questa analisi vengono scelte le domande pubblicate su Stack Overflow nel 2022 riguardanti i principali linguaggi di programmazione (secondo la classifica Stack Overflow Developer Survey 2022), si tratta di 827.141 istanze. Questo perché Stack Overflow è uno dei più grandi, se non il più grande, siti di domande e risposte riguardanti i linguaggi di programmazione e le tecnologie ad essi associate. In particolare, ogni domanda ha uno o più tag che indicano quale è il linguaggio di programmazione oggetto della domanda. Ad ogni domanda è stata attribuita un'emozione facendo riferimento alla classificazione delle emozioni proposta da Parrott nel 2001. È stata scelta la classificazione di Parrott perché coinvolge più di cento emozioni che vengono mappate in sei emozioni base: *love*, *joy*, *surprise*, *fear*, *sadness*, *anger*. A questa classificazione ne viene aggiunta una settima: *neutral,* per rappresentare la non espressione di emozioni.
La classificazione delle domande viene effettuata tramite un modello di deep learning basato su BERT, appositamente addestrato per classificare il testo analizzato secondo la variante dello schema di Parrott introdotta precedentemente. Il modello è stato addestrato e valutato su un dataset di 14.182 istanze. Il dataset è stato diviso in tre parti: training set (60%), evaluation set (20%) e test set (20%). Il training set è stato utilizzato per addestrare il modello mentre evaluation set è stato utilizzato sempre in fase di addestramento per verificare che il modello stesse generalizzando bene, quindi per evitare l'overfitting. Invece il test è stato utilizzato per misurare le prestazioni del modello. Il modello ha un'accuratezza del 66% nello stabilire a quale tra le sette classi appartiene una domanda, mentre ha un'accuratezza del 77% nello stabilire la polarità della domanda, ovvero nello stabilire se la domanda appartiene ad un'emozione positiva (*love*, *joy*, *surprise*) o negativa (*fear*, *sadness*, *anger*).
Per addestrare il modello e valutarne le prestazioni è stato necessario etichettare 14.182 istanze. Trattandosi di un'operazione molto dispendiosa se effettuata manualmente, l'etichettatura è stata ottenuta con una procedura automatica di distant supervision. In particolare, se individuata nel testo un'emozioni dello schema di Parrott questa viene fatta risalire all'emozione base che definirà l'etichetta. Invece se non viene identificata alcuna emozione viene attribuita un'etichetta neutrale.

Una volta classificate tutte le domande con il modello predittivo, sono state utilizzate per osservare l'andamento delle sette emozioni verso i linguaggi di programmazione nel corso dell'anno 2022.
La classificazione delle domande è stata anche utilizzata per realizzare una classifica dei linguaggi più apprezzati e meno apprezzati per l'intero ano 2022. I risultati di questa classifica sono stati confrontati con la sezione "Loved vs. Dreaded" del sondaggio Stack Overflow Developer Survey 2022, in cui sono stati intervistati più di 70.000 sviluppatori, ottenendo una bassa correlazione. Probabilmente perché le domande sono solitamente pubblicate da persone nuove nella programmazione mentre nel sondaggio citato sono stati intervistati sviluppatori tipicamente esperti.

Lo schema seguente mostra la variante del modello di Parrott utilizzata.



Lo schema seguente mostra l'andamento delle sei emozioni significative nel corso del 2022 per Python. Si può notare che l'andamento è praticamente costante eccetto in alcuni intervalli in cui si ha una brusca discesa e una successiva risalita si tratta di un periodo tipicamente scelto per le vacanze, motivo per cui calano le istanze.



La lista seguente mostra i linguaggi di programmazione e le tecnologie ad essi associate più apprezzate secondo il modello, con almeno mille istanze tra emozioni negative e positive. La lista è ordinata dalla più alla meno apprezzata.

MATLAB, R, SQL, Python, Assembly, C, Shell, Bash, Scala, CSS, VBA, C++, HTML, PowerShell, Go, JavaScript, Rust, PHP, Ruby, Java, TypeScript, C#, Swift, Kotlin, Dart, Solidity.