Progetto di Social Computing

RELAZIONE SULLA SECONDA PARTE

Borsoi Allison
mat. 147566
borsoi.allison@spes.uniud.it

Bosoppi Stefano
mat. 153845
bosoppi.stefano@spes.uniud.it

Canciani Mauro
mat. 150260
canciani.mauro@spes.uniud.it

a.a. 2022-2023

1 Introduzione

In questa relazione mostreremo gli elementi significativi della progettazione del nostro task, disponibile visitando il sito `https://sansbold-deploy-sc-2.s3.eu-south-1.amazonaws.com/secondo_progetto_social_computing/statement_truthfulness/index.html`.
In particolare vedremo se le precauzioni da noi prese per assicurare la qualità del lavoro si sono dimostrate sufficienti e, attraverso quelli che potrebbero essere considerati "metadati" (ad esempio il tempo medio impiegato per valutare un documento), saremo in grado di formulare alcune ipotesi sul comportamento che i lavoratori hanno assunto svolgendo il task.

2 Progettazione del task

Chiamiamo *task*, in italiano *compito*, l'aggregazione di piccoli compiti da far svolgere ad un lavoratore, ossia ad una persona che decide di svolgerli, al fine di raggiungere uno o più particolari obiettivi. Ciascuno di questi piccoli incarichi è poi ulteriormente suddiviso in *microtask*, ovvero dei lavori elementari, come può essere la risposta ad una domanda a scelta multipla.
Il task in questione è strutturato in tre compiti analoghi: la valutazione di un'affermazione e dell'associato giudizio di verità. I microtask di ognuno di questi invece consistono nella valutazione, nell'ordine, dell'affermazione sulla base della giustificazione del giudizio fornito; della qualità di tale giustificazione e, infine, della verità dell'affermazione, fornita con l'ausilio di una ricerca sulla rete.
In questa sezione tratteremo le strategie impiegate per garantire la qualità dei risultati prodotti dai lavoratori nello svolgimento di questo task.

2.1 Assegnazione dei documenti

Consideriamo come documento D una tripla (S, J, E), dove S è un'affermazione, J un giudizio (*supports* o *refutes*) ed E è la motivazione del giudizio. Vi sono due tipologie di documento: i *documenti di verifica*, G, ed i *documenti di lavoro*, N. I primi sono dei documenti che permettono di eseguire dei test nascosti, in modo da valutare la qualità del lavoro dell'utente, mentre i secondi sono documenti che permettono di ottenere dei risultati utili allo scopo del task, attraverso l'attività del lavoratore.
I documenti forniti sono stati ottenuti dalla fusione dei dataframe FEVER[1], per le affermazioni, ed e-FEVER[2], per le giustificazioni. Li abbiamo processati in modo da formare sei documenti del tipo (S_i, J_i^j, E_i^j), dove S_i è l'i-esima affermazione, J_i^j è il giudizio dell'i-esima affermazione fornita dal "punto di vista" j (ossia se fornito dal modello di machine learning o da un essere umano) e E_i^j è la spiegazione associata al giudizio.
Abbiamo considerato i documenti contenenti i giudizi, e le relative giustificazioni, fornite da un uomo come documenti di verifica. Da questa scelta consegue il fatto che abbiamo a disposizione tre documenti di verifica e tre documenti di lavoro.
Ogni task presenta quindi tre documenti, due di verifica ed uno di lavoro, come mostrato nella tabella 1. Abbiamo scelto questa particolare distribuzione dei documenti in modo da garantire che in ogni task vi siano sempre un documento di lavoro ed uno di verità che condividono la propria affermazione, al fine di verificare la coerenza di ogni lavoratore. In aggiunta, l'ordine di apparizione dei documenti in un task è pseudo-casuale per la libreria `random`[3] di Python.

Task	Documenti		
0	G_0	N_0	N_1
⋮	⋮	⋮	⋮
i	N_i	G_j	N_j
⋮	⋮	⋮	⋮
11	N_k	N_{11}	G_k

Tabella 1: Schema della distribuzione dei documenti tra i task

Sono poi evidenti due aspetti: il primo è che inevitabilmente vi sono dei task che presentano al loro interno documenti ripetuti; il secondo è che i requisiti stabiliti dalla consegna sono rispettati perché, per quanto stabilito prima, ogni documento di testo può essere valutato almeno tre volte da lavoratori differenti ed ogni documento di verifica può essere sottoposto ad utenti diversi almeno due volte.

2.2 Istruzioni significative

Per prendere ulteriori precauzioni in modo sul fatto che i lavoratori a cui avremmo sottoposto i vari task producessero lavori di qualità, è stato necessario curare attentamente le spiegazioni che sarebbero state fornite loro. Tra queste, abbiamo ritenuto opportuno spiegarne l'inserimento di due indicazioni: l'indipendenza di ogni elemento e le istruzioni sulla produzione di annotazioni.
L'istruzione che indirizzava l'utente a valutare ogni documento in maniera indipendente l'uno dall'altro[1] è stata necessaria per via della struttura vista al paragrafo 2.1: un lavoratore, infatti, deve valutare la stessa affermazione due volte. Se basasse il giudizio della seconda apparizione su quello che ha già fornito per la prima, il suo operato sarebbe stato condizionato, abbassando la qualità generale del lavoro.
Abbiamo ritenuto opportuno inserire anche l'istruzione di guida sulla produzione dell'annotazione[2] in quanto, nei test del sistema da noi eseguiti, sia l'azione di evidenziare il testo, che i criteri da applicare per evidenziarlo, non sembravano abbastanza intuitivi. Ciò, secondo noi, avrebbe potuto produrre tempi di compilazione più lunghi, arrecare maggiore stress all'utente, portando inevitabilmente ad una degradazione della bontà del suo lavoro. Quindi, sia per facilitare l'utilizzo del sistema per l'utente, sia per proteggere la qualità dei risultati, abbiamo concluso che fosse appropriato aggiungere questa istruzione.
In aggiunta, credendo nel fatto che una maggiore usabilità e chiarezza del sistema avrebbe facilitato la produzione di lavori di qualità, abbiamo riportato, per ogni microtask inerente ad un documento un riassunto, una riformulazione od un'applicazione delle istruzioni che regolavano direttamente, o indirettamente, lo svolgimento dello specifico microtask.

2.3 Vincoli del task

Un ulteriore stratagemma per garantire lavori di alto livello è stata la scelta di adeguati vincoli sul numero massimo di tentativi e sul tempo minimo per valutare ciascun documento.
Abbiamo scelto 3 come numero massimo di tentativi in quanto abbiamo voluto tutelare la possibilità di un lavoratore di fare il task da possibili problemi di connessione o, seppur ben più sporadici, errori da parte del sistema. Consideriamo questo valore come appropriato in quanto riteniamo probabile che, dopo il terzo tentativo di compilazione, lo stesso lavoratore perda la voglia di compilare il task.
Per quanto riguarda il tempo minimo imposto da trascorrere per valutare ciascun documento, ci è sembrato ragionevole fissarlo a due minuti. Questo valore è il risultato di alcuni test sul funzionamento del sistema da noi svolti, aventi l'obiettivo di verificare quanto sarebbe potuta durare una compilazione veloce, ma comunque adeguata, del task.

3 Analisi dei risultati

In questa sezione analizzeremo la qualità dei dati forniti dai lavoratori che hanno svolto il nostro task, a cui hanno potuto accedere dalla piattaforma *Amazon Mechanical Turk* seguendo il link incluso nell'*H.I.T.*, `https://sansbold-deploy-sc-2.s3.eu-south-1.amazonaws.com/secondo_progetto_social_computing/statement_truthfulness/index.html`.
Nell'analizzare i risultati ottenuti, il nostro interesse non sarà limitato a constatare se le risposte fornite da un singolo lavoratore sono effettivamente corrette, ma cercare di capire come l'intero insieme di utenti si è comportato nel rispondere al sondaggio.

[1] L'istruzione definita come "di indipendenza" è qui riportata: *Evaluate each statement independently. If you are asked to judge the same statement again, do not base your evaluation on the one you have already given.*

[2] Il testo dell'istruzione è il seguente: *When you first evaluate a statement, you can only base your opinion of its truthfulness on the explanation you are given. Highlight in the explanation which words/sentences help you make up your judgment and then press the Select button.*

3.1 Qualità dei worker

Una misura che può permettere di inferire la qualità di un lavoratore rispetto agli altri è il *percent agreement*. Nel nostro caso, abbiamo calcolato l'accordo percentuale tra una coppia di lavoratori, per una dimensione di valutazione fissata, nel seguente modo:

$$p(w_i, w_j) = \frac{E_{i,j}}{T_{i,j}} \cdot 100$$

dove w_i e w_j sono due lavoratori qualsiasi, $E_{i,j}$ è il numero di volte in cui w_i e w_j rispondono allo stesso modo, nello stesso microtask riferito allo stesso documento e $T_{i,j}$ è il numero di volte in cui w_i e w_j valutano lo stesso documento. Nel caso in cui w_i e w_j non abbiano nessun documento in comune, consideriamo l'accordo percentuale come "indefinito" e quindi $p(w_i, w_j) = -1$.
È facile notare che, nei casi in cui il percent agreement è definito, questa misura d'accordo varia nell'intervallo $[0, 100]$, dove 0 rappresenta totale disaccordo tra la coppia di lavoratori e, simmetricamente, 100 indica totale accordo tra le parti.

	A	**B**	**C**	**D**	**E**	**F**	G	**H**	I	J
A	100.0	0.0	0.0	0.0	0.0	0.0	-1.0	0.0	50.0	0.0
B	0.0	100.0	0.0	0.0	0.0	100.0	-1.0	0.0	33.3	0.0
C	0.0	0.0	100.0	100.0	0.0	0.0	100.0	0.0	0.0	0.0
D	0.0	0.0	100.0	100.0	100.0	100.0	-1.0	100.0	0.0	50.0
E	0.0	0.0	0.0	100.0	100.0	66.7	-1.0	66.7	0.0	100.0
F	0.0	100.0	0.0	100.0	66.7	100.0	-1.0	33.3	100.0	100.0
G	-1.0	-1.0	100.0	-1.0	-1.0	-1.0	100.0	-1.0	-1.0	-1.0
H	0.0	0.0	0.0	100.0	66.7	33.3	-1.0	100.0	0.0	50.0
I	50.0	33.3	0.0	0.0	0.0	100.0	-1.0	0.0	100.0	0.0
J	0.0	0.0	0.0	50.0	100.0	100.0	-1.0	50.0	0.0	100.0

Tabella 2: Percent agreement per la dimensione `truthfulness_1`, il giudizio di verità dell'utente basato sul giudizio e sulla spiegazione forniti.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
A	100.0	0.0	100.0	66.7	100.0	100.0	-1.0	100.0	0.0	100.0
B	0.0	100.0	0.0	50.0	0.0	0.0	-1.0	0.0	66.7	0.0
C	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
D	66.7	50.0	0.0	100.0	100.0	100.0	-1.0	100.0	50.0	50.0
E	100.0	0.0	0.0	100.0	100.0	66.7	-1.0	66.7	0.0	100.0
F	100.0	0.0	0.0	100.0	66.7	100.0	-1.0	33.3	0.0	100.0
G	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	100.0	-1.0	-1.0	-1.0
H	100.0	0.0	0.0	100.0	66.7	33.3	-1.0	100.0	0.0	50.0
I	0.0	66.7	0.0	50.0	0.0	0.0	-1.0	0.0	100.0	0.0
J	100.0	0.0	100.0	50.0	100.0	100.0	-1.0	50.0	0.0	100.0

Tabella 3: Percent agreement per la dimensione `truthfulness_2`, il giudizio di verità dell'utente basato sulla sua ricerca online.

Abbiamo quindi calcolato il percent agreement per tutte le coppie di lavoratori, per ambedue le dimensioni categoriali, cioè quelle che la consegna ha disposto di chiamare `truthfulness_1` e `truthfulness_2`. I risultati sono riportati nelle tabelle 2 e 3, dove abbiamo identificato ciascun worker con una lettera maiuscola dell'alfabeto. Per entrambe si può notare come abbiamo collezionato le risposte di solo dieci lavoratori. Approfondendo, abbiamo osservato come anche due task, da noi codificati con i numeri 1 e 2, non sono stati sottoposti ad alcun utente. Tuttavia, nei file con i dettagli del lavoratori forniti dal sistema, *Crowd_Frame*[4], erano presenti dodici voci. Confrontando quindi la lista di lavoratori completa con quella contenente le risposte memorizzate, abbiamo notato che l'unico dato che accomunava questi due utenti era il fatto che entrambi avevano compilato il sondaggio da un terminale con un indirizzo IPv6. Abbiamo quindi teorizzato, ma non abbiamo avuto l'occasione di verificarlo, che questo potesse essere dovuto ad un problema del sistema utilizzato.
Un altro fatto anomalo è osservabile dalla riga, e quindi dalla colonna, relativa al lavoratore G, visibile sia nella tabella 2 che nella 3. Si può notare infatti che solamente tre valori sono definiti: questo è dovuto al fatto che G ha completato solamente le microtask di un documento su tre. Non ci siamo sbilanciati nell'ipotizzare il perché

solamente parte del suo task è stato memorizzato, tuttavia è chiaro che lo stesso comportamento del lavoratore è stato anomalo, in quanto ha trascorso ben più di venticinque minuti sul suo primo documento[3].
Per quanto riguarda il resto dei lavoratori, si può notare che, in generale, sono presenti un numero molto alto di zeri nella tabella 2, indice di un disaccordo generale tra i lavoratori. Mentre, nella tabella 3 possiamo vedere aumentare il numero di 100, probabilmente per il fatto che, cercando le informazioni in rete, è possibile per i lavoratori convergere su un valore di verità che ciascuno di loro può ritenere "definitivo" e scarsamente soggetto ad interpretazione, in contrasto con quella che potrebbe essere la spiegazione del giudizio fornita nel documento.

3.2 Statistiche

Ulteriori dati che abbiamo calcolato sono la percentuale media di testo annotato, il numero di volte in cui gli utenti aggiornano la loro annotazione della spiegazione ed il tempo medio impiegato per ciascun documento. Riteniamo importanti queste osservazioni perché, come spiegheremo meglio in seguito, ci possono permettere di ricavare informazioni su come il lavoratore ha svolto il suo compito.

Spiegazione	Percentuale media annotata
Winter's Tale was released in 2014.	88.9
Hush (2016 film) was produced by Trevor Macy and Jason Blum.	59.1
Anne Rice was born in New Orleans, Louisiana, which is in the United States of America.	51.6
The claim is that Winter's Tale was released in 1987. The evidence states that Winter's Tale is a 1983 novel by Mark Helprin. This is a novel, so it wasn't released in 1987. Therefore, the claim is false.	32.0
The claim is that Anne Rice was born in the United States of America. The evidence states that she was born in New Orleans and that New Orleans is a major United States port. Therefore, the claim is true.	30.8
The evidence says that the film was produced by Trevor Macy and Jason Blum. The claim says that the film was produced by Jason Blum. So, the answer must be false because the film was produced by Trevor Macy, not just Jason Blum.	30.2

Tabella 4: Classifica, per spiegazione, della percentuale media di testo annotato

La percentuale media di testo annotato è utile per capire, in media, di quanta informazione, rispetto a quella fornita, ha bisogno l'utente per poter formulare un proprio giudizio di verità. Dalla tabella 4 possiamo intuire che tale quantità è "costante", perché è evidente che, all'aumentare della lunghezza della spiegazione, diminuisce la porzione di parole ritenute utili dall'utente al fine di giudicare l'affermazione a lui presentata.

Spiegazione	Numero di aggiornamenti
The claim is that Winter's Tale was released in 1987. The evidence states that Winter's Tale is a 1983 novel by Mark Helprin. This is a novel, so it wasn't released in 1987. Therefore, the claim is false.	2
The evidence says that the film was produced by Trevor Macy and Jason Blum. The claim says that the film was produced by Jason Blum. So, the answer must be false because the film was produced by Trevor Macy, not just Jason Blum.	2
The claim is that Anne Rice was born in the United States of America. The evidence states that she was born in New Orleans and that New Orleans is a major United States port. Therefore, the claim is true.	1

Tabella 5: Classifica delle spiegazioni maggiormente aggiornate, per numero di aggiornamenti

Il numero di aggiornamenti di un'annotazione sono il numero di volte in cui l'utente, avendo già selezionato una specifica parte della spiegazione, decide di modificare la sua scelta, selezionandone una nuova. Non viene contata, ovviamente, la prima selezione, in quanto nessuna frazione della spiegazione era già annotata.

[3]Questo tempo ci ha sorpreso perché, oltre ad essere decisamente superiore alle nostre stime, si discosta completamente dalle misurazioni effettuate nel paragrafo 3.2

Dalla tabella 5 possiamo osservare che solo le spiegazioni fornite dal modello di machine learning hanno subito aggiornamenti. Le possibili giustificazioni che possiamo dare a questo fenomeno sono due, una di natura tecnologica ed una di natura psicologica. La prima riguarda il fatto che, oggettivamente, è difficile eseguire la selezione su dispositivi mobili e spesso produce risultati che non soddisfano l'utente, costringendolo ad aggiornare la spiegazione, non per un cambio di idea, ma per necessità. Inoltre, con un testo più lungo, aumenta la probabilità di evidenziare erroneamente qualcosa che l'utente ritiene "improprio" o "sbagliato", portando al fenomeno che abbiamo osservato.
Per quanto riguarda la seconda ipotesi, quella di natura psicologica, è possibile che il lavoratore sia maggiormente indeciso su cosa evidenziare quando ha davanti a se un testo medio-lungo e quindi maggiormente soggetto a ripensamenti.
Queste ovviamente sono solamente speculazioni non verificate, ma è comunque innegabile che le spiegazioni di lunghezza maggiore sono anche quelle che hanno subito aggiornamenti.
Infine, nella tabella 6, possiamo osservare i tempi di svolgimento di ciascun documento. Ciò che balza subito all'occhio è che i documenti di lavoro hanno tempo di svolgimento maggiore rispetto a quelli di verifica: questa può essere considerata come una prova del fatto che per le spiegazioni più lunghe, mediamente, il lavoratore impiega più tempo, oltre che a leggere, a selezionare la porzione di testo che ritiene migliore.
In aggiunta, i risultati della tabella sembrano appoggiare la stima da noi fatta nel paragrafo 2.3: i tempi impiegati nel valutare ciascun documento superano i due minuti.

ID documento	Tempo (s)
N_166632	291
N_51526	162
N_77465	156
G_77465	146
G_166632	131
G_51526	124

Tabella 6: Classifica dei documenti su cui gli utenti hanno mediamente trascorso più tempo

Ci teniamo, infine, a precisare che per produrre i risultati mostrati nelle tabelle 4, 5 e 6, non abbiamo considerato il lavoro di G[4].

4 Conclusioni

In conclusione, abbiamo potuto osservare che le misure di precauzione messe in atto, almeno per quanto discusso al paragrafo 3.1, non sono risultate efficaci per garantire un accordo, ma hanno comunque permesso ai vari task di rispettare i requisiti di accesso stabiliti dalla consegna, altrimenti ci sarebbero state altre celle con valore indefinito nella tabelle 2 e 3.
Siamo però riusciti ad osservare, però, come un utente ha, mediamente, bisogno di una quantità di informazioni costante per poter esprimere un giudizio di verità di affermazione e che uno dei fattori per la sua incertezza potrebbe essere la lunghezza del testo informativo a lui fornito.
Siamo inoltre convinti che sarebbe necessaria una nuova iterazione dei task in modo da poter eseguire un analisi con dodici lavoratori di qualità, tuttavia è stato per noi evidente come non possiamo controllare il tipo di dispositivo con cui l'utente accede al task, si veda il problema degli indirizzi IPv6 nel paragrafo 3.1, ed il tempo massimo che impiega a rispondere ai microtask associati ad un documento, si è solo potuto configurare un tempo massimo per l'intero task attraverso la piattaforma *Amazon Mechanical Turk*.

Riferimenti bibliografici

[1] James Thorne et al. «FEVER: a Large-scale Dataset for Fact Extraction and VERification». In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, giu. 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074.

[2] Dominik Stammbach e Elliott Ash. «e-FEVER: Explanations and Summaries for Automated Fact Checking». en. In: *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*. A cura di Emiliano De Cristofaro e Preslav Nakov. Conference for Truth and Trust Online (TTO 2020) (virtual); Conference Location: online; Conference Date: October 16-17, 2020; Due to the Coronavirus (COVID-19) the conference was conducted virtually. Arlington, VA: Hacks Hackers, 2020, pp. 32–43. ISBN: 978-1-7359904-0-8. DOI: 10.3929/ethz-b-000453826.

[3] Python. *random — Generate pseudo-random numbers*. Ultima visita nel 13 gennaio 2023. 2021. URL: https://docs.python.org/3.10/library/random.html.

[4] Si vedano le tabelle 2 e 3

[4] Michael Soprano et al. ≪Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf≫. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM '22. Virtual Event, AZ, USA: Association for Computing Machinery, 2022, pp. 1605–1608. ISBN: 9781450391320. DOI: 10.1145/3488560.3502182. URL: https://doi.org/10.1145/3488560.3502182.