Progetto ReviewLand: Relazione sulle attività del primo anno del progetto (Settembre 2016 - Agosto 2017)

Dati generali

- Progetto ReviewLand, progetto co-finanziato dalla Fondazione Cassa di Risparmio di Lucca e da IIT-CNR.
- Consorzio: composto da 3 enti di ricerca e sviluppo: Istituto di Informatica e Telematica, CNR, Pisa, Italy, LUCENSE ScaRL, Lucca, Italy, e IMT Scuola Alti Studi, Lucca, Italy.
- Partners finanziati: IIT-CNR, Pisa, Italy e Lucense ScaRL, Lucca, Italy.
- Lettera di accettazione della Fondazione CRL datata 19 Agosto 2016 e protocollata da IIT-CNR in data 29 Agosto 2016, protocollo n. 07044.
- Data d'inizio: 1 Settembre 2016.
- Durata prevista: 18 mesi.
- Importo Complessivo Finanziato: Euro 40,000 (20,000 per l'anno 2016, 20,000 per l'anno 2017).
- Tematiche: il progetto ReviewLand si occupa di studiare e realizzare metodologie per l'estrazione, l'elaborazione e l'interpretazione di recensioni online, reperibili dai principali siti di "e-advice". L'indagine si concentra su dati che hanno come oggetto il territorio di Lucca e provincia.
- Sito web: reviewland.projects.iit.cnr.it
- Coordinatore: Marinella Petrocchi, Trustworthy and Secure Future Internet Group, IIT-CNR, Pisa.
- Team: Il gruppo di ricerca è formato da Marinella Petrocchi, Angelo Spognardi, attualmente ricercatore presso l'università Sapienza di Roma e DTU Compute, Denmark, Michela Fazzolari e Massimiliano La Gala, assegnisti IIT finanziati dal progetto, Alessandro Tommasi e Cesare Zavattari, collaboratori di LUCENSE ScaRL.

Scopo: Internet in generale e i Social Media in particolare offrono una grande mole di commenti e opinioni che vengono giornalmente inseriti da milioni di utenti. Questi testi riguardano i temi più disparati: apprezzamenti o critiche a politici e celebrità, commenti su eventi pubblici, suggerimenti per acquisti e/o viaggi.

Sia i fornitori di servizi e prodotti, sia i consumatori, possono trarre vantaggio dall'estrazione di informazioni rilevanti, utili e veritiere, da questa miriade di post, tweet e commenti disponibili online. I primi, ad esempio, ne beneficiano per correggere le loro campagne pubblicitarie e le loro linee di prodotti. I secondi ottengono informazioni utili su un determinato bene o servizio di interesse. In particolare, le recensioni online determinano la qualità percepita di un prodotto/servizio e guidano l'utente nella scelta.

Lo scopo di Reviewland è quello di studiare, innovare e realizzare ex-novo, metodologie per l'estrazione, l'elaborazione, e l'interpretazione di recensioni online. L'indagine si concentra sulle recensioni esistenti sui principali siti di e-advice (come ad esempio Booking e Tripadvisor), prendendo in esame i dati disponibili per il territorio di Lucca e provincia.

ReviewLand si sviluppa su un periodo temporale di 18 mesi e si articola nelle seguenti principali attività:

1. Estrazione ed acquisizione dei dati da piattaforme di online reviews; "data cleaning", arricchimento semantico dei dati e memorizzazione su database strutturato;
2. Elaborazione e analisi dei dati (e.g., individuazione di trend, caratterizzazione dei recensori, discrepanze testo-voto);
3. Divulgazione dei risultati

Risultati: Di seguito, si elencano i risultati scientifici ottenuti dal progetto nel suo primo anno di attività, sul periodo Settembre 2016 - Agosto 2017.

1. *Studio dello stato dell'arte e acquisizione dei dati*: il primo periodo del progetto è stato dedicato allo studio delle tecniche esistenti nell'ambito del trattamento delle informazioni online. Sono stati realizzati degli strumenti informatici (*crawler*) per acquisire le informazioni presenti su due popolari siti di e-advice (Booking.com e TripAdvisor.com). Sono state quindi raccolte da questi due siti

più di 1 milione di recensioni online, insieme ai corrispondenti meta-dati (titolo, data, recensore, categoria del recensore, ect.). Per effettuare uno studio in grado di generalizzare i risultati ottenuti su ampia scala, oltre ai dati concernenti la citta di Lucca, le recensioni acquisite da Booking.com sono anche relative a tutte le strutture ricettive presenti in 10 metropoli, nei 5 continenti. Per quanto riguarda TripAdvisor, le recensioni provengono da tutte le strutture ricettive, ristoranti e attività che riguardano sia la città che la provincia di Lucca. Sono state inoltre acquisite tutte le recensioni effettuate in luoghi diversi da Lucca, dai recensori che hanno recensito almeno una volta Lucca. Tutte le informazioni raccolte sono state sottoposte a un processo di *data cleaning*, in modo da migliorare la qualità dei dati acquisiti e ridurne il "rumore". I dati sono stati in seguito inseriti in una base di dati strutturata. I dati sono utilizzati esclusivamente per le finalità del progetto, nel rispetto delle normative vigenti sulla tutela della privacy.

2. *Elaborazione e analisi dei dati*: i dati raccolti sono stati utilizzati per effettuare varie analisi.
   - *Discrepanze testo-voto di recensioni online:* Un'analisi è stata condotta per identificare discrepanze tra il testo di una recensione e il voto ad essa associato, con particolare riferimento alle recensioni di hotel. La maggior parte dei siti di e-commerce mostra per ogni prodotto una lista di recensioni, ciascuna delle quali è associata ad un voto numerico. La valutazione complessiva del prodotto è calcolata in base ai singoli voti associati alle recensioni e spesso riassunta in un voto numerico da 1 a 5. L'assunzione implicita di questo processo è che esista una corrispondenza certa tra il testo di una recensione ed il voto numerico ad essa associato. In questo ambito, è stata condotta una analisi volta a identificare la presenza di possibili discrepanze (mismatches) nella corrispondenza tra testo e voto di una recensione. Di fatto, un contenuto testuale è soggettivo, perciò testi che esprimono opinioni diverse possono presentare lo stesso voto e, viceversa, a testi con opinioni simili può essere assegnato un voto diverso. L'identificazione di recensioni che presentano un disaccordo tra testo e voto è utile sia per i fornitori di servizi e prodotti, sia per gli utenti. Questo studio è stato realizzato sulle recensioni di hotel, raccolte sia da Booking che da Tripadvisor, utilizzando un classificatore di testo addestrato per identificare recensioni con discrepanza tra testo e voto. I risultati di questo studio sono stati riportati in un articolo che è stato pubblicato in una rivista ISI di livello internazionale (Cognitive Computation, Springer) [1].
   - *Clustering di recensioni online* Una seconda analisi ha riguardato il raggruppamento di recensioni online. L'analisi è avvenuta in collaborazione con LUCENSE. Il tema è stato affrontato in passi successivi: a) è stata studiata e definita una nuova metrica chiamata *aderenza*, in grado di misurare il livello di corrispondenza di un testo rispetto ad una terminologia tipica relativa al dominio di appartenenza del testo stesso; b) successivamente, è stata studiata la correlazione tra aderenza e voto assegnato alle recensioni. Utilizzando la metrica proposta, sono stati eseguiti vari tipi di esperimenti per verificare quanto la nuova metrica fosse informativa del voto assegnato ad una recensione. Tale correlazione è evidente qualora si considerino non singole recensioni, ma gruppi di recensioni, di fatto si è osservato che l'aderenza media di un insieme di recensioni è proporzionale al voto medio, quando si considerano gruppi non adiacenti; c) è stato infine proposto un meccanismo per raggruppare le recensioni che sfruttasse la metrica proposta anziché il voto. I vantaggi di questo approccio sono molteplici: l'aderenza può essere calcolata su testi anche quando non è disponibile un voto associato e non è richiesta alcuna conoscenza a priori del dominio di riferimento o della lingua in cui i testi sono stati scritti. Inoltre il meccanismo proposto non richiede l'utilizzo di un insieme di recensioni etichettate per l'addestramento (approccio non-supervisionato). La metrica proposta è stata utilizzata per raggruppare recensioni con opinioni simili. Infatti si è osservato che dividendo le recensioni in gruppi ordinati secondo l'aderenza media, il gruppo con aderenza maggiore mostra anche un voto medio maggiore rispetto al gruppo con aderenza media minore (e voto medio minore). I risultati di questa ricerca sono stati riassunti in un lavoro [2] pubblicato negli atti di una conferenza di livello internazionale (17th International Conference on Web Engineering, ICWE 2017).
   - *Analisi strettamente relative a Lucca e provincia*: Una terza analisi riguarda da vicino le recensioni disponibili per strutture relative a Lucca e alla provincia di Lucca. Inizialmente è stato effettuato uno studio statistico sui dati per estrarre informazioni rilevanti sull'afflusso dei visitatori. A tale scopo è stata analizzata la distribuzione geografica dei visitatori di Lucca, che hanno lasciato almeno una recensione sul sito di TripAdvisor, che ha confermato l'interesse internazionale nei confronti della città. In seguito, una analisi temporale dei voti delle recensioni ha permesso di rilevare relazioni interessanti tra la qualità ricettiva percepita dai

visitatori e il periodo dell'anno in cui avviene la visita. Inoltre, grazie alla disponibilità di dati provenienti da due diverse piattaforme di e-advice, (Booking e Tripadvisor) è stato possibile effettuare una analisi comparativa tra i due datasets, che ha confermato la validità generale delle conclusioni ottenute. Infine, è stata realizzata una analisi sulle preferenze di viaggio dei visitatori di Lucca, in base alla nazionalità di provenienza, estraendo sia patterns di visite nel territorio di Lucca e provincia, sia su scala internazionale. Le analisi su tali "tour virtuali" (ovvero, tour dei visitatori di Lucca, ottenibili a partire dai dati disponibili online) saranno approfonditi nella seconda e ultima fase del progetto. I primi risultati di questa ricerca sono riassunti in un articolo attualmente sottomesso ad una rivista di livello internazionale [3].

Sviluppi futuri del progetto: La terza attività del progetto riguarda la divulgazione dei risultati ottenuti. A questo proposito, oltre alle pubblicazioni scientifiche derivate dall'attività di ricerca, il team di lavoro ha in programma la partecipazione alla manifestazione "Bright 2017 - La notte dei ricercatori in Toscana", con stand e poster presso IMT Lucca, il giorno 29 Settembre 2017. Allestiremo lo stand e prepareremo il poster, con lo scopo di mostrare al pubblico i risultati ottenuti finora.

Nell'ambito dell'attività di ricerca relativa al progetto, è in corso un ulteriore - e particolarmente ambizioso - studio, in collaborazione con Lucense. Lo studio si prefigge di verificare la possibilità di generare recensioni online in maniera completamente automatica, e tali da essere considerate recensioni genuine (e scritte da umani). Attualmente, è in corso la fase di applicazione di tecniche di intelligenza artificiale basate su *deep learning*, per la generazione automatica delle recensioni. Nei mesi successivi, seguirà la fase di validazione delle recensioni cosi' generate.

Inoltre, una ulteriore estensione delle analisi relative a Lucca e provincia è attualmente in fase di studio con lo scopo di valorizzare ulteriormente le recensioni raccolte. L'obiettivo sarà quello di scoprire percorsi ricorrenti dei visitatori della città e della provincia di Lucca.

In collaborazione con IMT Lucca, stiamo conducendo una analisi di gradimento sulla popolare manifestazione Lucca Comics & Games. Stiamo analizzando i tweets relativi all'ultima manifestazione (Novembre 2016), allo scopo di valutare le opinioni dei visitatori, sia riguardo all'evento in sé, sia riguardo alla città.

Infine, con particolare riferimento alle recensioni sui ristoranti di Lucca e provincia, e' in corso d'opera un lavoro in collaborazione con l'Università della Calabria, per verificare la veridicità e l'attendibilità delle recensioni stesse, investigando così la presenza dell'impellente fenomeno delle *fake reviews* in recensioni riguardanti Lucca.

Sito web: Sin dall'inizio del progetto, è stato realizzato un sito web per promuovere i risultati ottenuti, che viene costantemente aggiornato. Il sito è raggiungibile al seguente indirizzo: *reviewland.projects.iit.cnr.it* Nel sito è possibile trovare una breve descrizione del progetto, una sezione contenente le pubblicazioni prodotte, il link a materiale che descrive i risultati di ricerca ottenuti finora, il team di lavoro e altre informazioni aggiuntive.

Assegnisti co-finanziati dal progetto:

- Massimiliano La Gala, Bando IIT 17/2016 (protocollo bando IIT 7957, del 30/09/2016).
- Michela Fazzolari, Bando IIT 3/2016 (protocollo bando IIT 0001022, del 15/02/2016).

Pubblicazioni

1 Michela Fazzolari, Vittoria Cozza, Marinella Petrocchi, Angelo Spognardi. A study on text-score disagreement in online reviews. Springer Cognitive Computation, accepted, 2017. Si allega il pdf del lavoro, Allegato A.

2 Michela Fazzolari, Marinella Petrocchi, Alessandro Tommasi, Cesare Zavattari. Mining worse and better opinions: Unsupervised and agnostic clustering of online reviews. In Proceedings of 17th International Conference on Web Engineering, pp.494-506, Springer (2017). Si allega il pdf del lavoro, Allegato B.

3 Michela Fazzolari, Massimiliano La Gala, Marinella Petrocchi. Enriching Traveling Information through Online Reviews Analysis. Submitted, 2017. Si allega il pdf del lavoro, Allegato C.

Alla presente relazione tecnica, si allegano inoltre:

- la prima relazione delle attività di Lucense ScaRL (Allegato D);

- la rendicontazione finanziaria del progetto.

Pisa, 15 Settembre 2017 In fede, Marinella Petrocchi