Nome paper: Predicting Personality from Twitter - Golbeck, Robles, Edmondson, Turner

Obiettivo: Predire la personalità di un utente attraverso le informazioni pubbliche del suo profilo Twitter ed evitando test della personalità

Dataset usato: Hanno creato un'applicazione per Twitter con 2 funzioni: somministrare un Big5P Test (45 domande) a 50 utenti e raccogliere gli ultimi 2000 tweet di ciascuno.

Feature usate: #followers, #following, #menzioni* (qualcuno ti cita con @username in un tweet), densità rete sociale (non dice come è stata calcolata), #risposte*, #hashtag*, #link*, parole per tweet

*sia il raw number sia la media per tweet

Tecniche usate: Hanno messo tutti i tweet di un utente in un unico documento. Fase 1: hanno provato a cercare delle correlazioni tra il contenuto dei documenti (feature analizzate dal punto di vista psico-linguistico) e i tratti del B5PTest (hanno ottenuto poche correlazioni). Fase 2: hanno provato a predire i tratti del B5Ptest "using the profile data as a feature set" con l'ausilio di 2 algoritmi di machine learning (risultato: predizione discostanti massimo dell'11%-18% dei valori attuali raccolti col questionario).

Fase 1.

Su ogni documento contenente i tweet degli utenti hanno fatto lavorare 3 tool: LIWC, MRC Psycholinguistic Database ed il General Inquirer dataset.

- LIWC produces statistics on 81 different features of text in five categories. These include Standard Counts (word count, word per sentence), Psychological Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (such as occupation, financial issues, health), and Other dimensions (swear words). Correlations between these features and personality traits (e.g. anxiety words and neuroticism scores) would not be surprising. This produced 79 text features.
- MRC Psycholinguistic Database is a list of over 150,000 words with linguistic and psycholinguistic features of each word. We computed the average non-zero score for each feature over all the words from each user.
- Using the General Inquirer dataset we performed a word by word sentiment analysis of each user's tweets. GID provides a hand annotated dictionary that assigns words sentiment values on a -1 to +1 scale. We computed a score for each user that was the average sentiment score for all words used in their list of tweets

We began by running a Pearson correlation analysis between subjects' personality scores and each of the features obtained from analyzing their tweets and public account data. Risultati in Allegato1.

Fase 2.

To predict the score of a given personality feature, we performed a regression analysis in Weka. We used two regression algorithms: Gaussian Process and ZeroR, each with a 10-fold cross-validation with 10 iterations. Risultati in Allegato2.

Risultati ottenuti: Già anticipate prima. "In our previous work studying personality on Facebook (dopo c'è l'analisi del paper), we had fewer and weaker correlations, but were able to predict all personality traits to within roughly 11%. This analysis of Twitter data yielded similar results for openness and agreeableness, but less impressive results for conscientious, extraversion, and neuroticism."

Nome paper: Predicting personality with social media - Golbeck, Robles, Turner

Obiettivo: Predire la personalità di un utente attraverso le informazioni pubbliche del suo profilo Facebook ed evitando test della personalità

Dataset usato: Hanno creato un'applicazione per Facebook con 2 funzioni: somministrare un Big5P Test (45 domande) a 279 utenti e raccogliere le informazioni pubbliche di ciascuno.

Feature usate: densità del grafo sociale ("i.e. what percentage of possible edges between friends exist."); tutte le informazioni personali possibili ricavabili da Facebook (nome, compleanno, stato sentimentale, religione, istruzione, sesso, città natale); statistiche interne (data iscrizione a Facebook, se utilizza app, numero di note); informazioni aggiuntive derivate (numero di esperienze descritte, se ha specificato la città natale o la religione); numero di caratteri nella voci in cui l'utente descrive le sue cose preferite o le sue attività (più scrivi più sei Open); stati (con rispettive lunghezze); campo "About me" e "blurb text" proveniente dal profilo (combinato in un'unica stringa – di lunghezza media 26 parole)

Tecniche usate: Hanno scartato gli utenti con 10 o meno parole nella stringa: sono rimasti 167 utenti con 42 parole in media) – dicono che comunque non ci sarebbero stati cambiamenti significativi considerando anche gli utenti con poche parole.

Con LIWC hanno analizzato il testo e prodotto 81 feature in 5 categorie (come prima). Come prima hanno calcolato i coefficienti di relazione di Pearson tra queste feature e i tratti del B5PTest. Risultati in Allegato3 (deboli correlazioni). Come nel precedente paper, alcune feature sono correlate a tratti del carattere (Consciousness -> poche parolacce negli stati; Extroverts -> molti amici, rete sparsa e ricche descrizioni delle attività passate; Donne -> più Neurotic (non serviva uno studio per scoprirlo); Agreeableness -> affective process words).

Per predire la personalità hanno usato le feature significative dal profilo e le feature linguistiche (escludendo feature con stesso valore per ogni utente).

Hanno una regressione lineare multipla per ogni fattore della personalità, producendo un vettore di pesi per ogni feature.

Per predire lo score di un dato fattore per un utente hanno eseguito una regressione lineare in Weka con una 10-fold cross validation con 10 iterazioni, utilizzando 2 algoritmi: M5' Rule e Gaussian Processes. Risultati in Allegato4. MAE: 11%

Risultati ottenuti: Sono riusciti, con le informazioni dei profili, a predire i tratti del B5PTest con un discostamento massimo dell'11% dai valori attuali raccolti coi questionari.

Nome paper: Our Twitter Profiles, Our Selves: Predicting Personality with Twitter – Quercia, Kosinski, Stillwell, Crowcroft

Obiettivo: Studiare i tratti del B5PTest per 5 categorie di utenti di Twitter (altri studi erano già stati fatti su Facebook): listeners (chi segue molte persone), popular (chi è seguito da molti), highly-read (chi è spesso elencato nel reading lists degli altri), influential (secondo il punteggio di Klout), influential (secondo la misura utilizzata da TIME: (2*#followersTwitter + #amiciFacebook)/2).

Dataset usato: Con l'applicazione myPersonality su Facebook hanno fatto compilare il test della personalità a 5 milioni di utenti. Il 40% ha acconsentito a condividere il risultato del test, ma solo poche centinaia lo hanno postato su Twitter. Hanno perciò considerato solo i 335 utenti di Facebook che avevano specificato anche un profilo Twitter.

Feature usate: Hanno considerato le seguenti 5 caratteristiche: log(#followers), log(#following), log(#listing), Klout influence, TIME influence.

"We are interested in logarithm because the corresponding distributions are not normal and their logarithm transformations account for the violation of normality". Non sono sicuro di aver compreso le ragioni di questa scelta.

Tecniche usate: Hanno calcolato il "Pearson product-moment correlation" (credo sia un altro modo per definire l'indice di correlazione di Pearson, o sbaglio?) tra ciascuna delle 5 caratteristiche "OCEAN" e le 5 feature elencate in precedenza. Risultati in Allegato5. Hanno dato un'interpretazione dei risultati, come nei 2 paper precedenti.

Hanno predetto la personalità per utenti di Twitter considerando solo la struttura del grafo sociale (#followers/following/listed). Quindi questa analisi può essere condotto anche su profili non pubblici che nascondo i tweet (poiché quelle informazioni sono comunque disponibili pubblicamente).

Hanno eseguito una regressione lineare con una 10-fold cross validation con 10 iterazioni, utilizzando l'M5' Rules. Hanno misurato l'RMSE tra valori predetti e valori attuali. L'RMSE massimo è stato 0.88 sull'Extroversion (da notare che, in tutti i paper, l'Openness è quella più facilmente predicibile): dicono che il risultato è ottimo in quanto il Netflix prize è andato a un team avente RMSE di 0.8567.

Risultati ottenuti: Influentias sono Estroversi, Conscious e poco Neurotici; i popolari hanno alta Openness; ecc. Vedere Allegato6.

Nome paper: Private traits and attributes are predictable from digital records of human behavior – Kosinski, Stillwell, Graepel

Obiettivo: Dimostrare l'accuratezza con la quale è possibile predire, attraverso record del comportamento umano (tweet, Facebook like, ecc.), tratti della personalità che un utente riteneva essere riuscito a mantenere privati (sesso, religione, orientamento politico, ecc). Nel paper vengono utilizzati i like su Facebook (foto, stati, pagine)

Feature usate: We selected traits and attributes that reveal how accurate and potentially intrusive such a predictive analysis can be, including "sexual orientation," "ethnic origin," "political views," "religion," "personality," "intelligence," "satisfaction with life" (SWL), substance use ("alcohol," "drugs," "cigarettes"), "whether an individual's parents stayed together until the individual was 21 y old," and basic demographic attributes such as "age," "gender," "relationship status," and "size and density of the friendship network."

Dataset usato: The study is based on a sample of 58,466 volunteers from the United States, obtained through the myPersonality Facebook application, which included their Facebook profile information, a list of their Likes (n = 170 Likes per person on average), psychometric test scores, and survey information. Five Factor Model personality scores were established using the International Personality Item Pool questionnaire with 20 items. Intelligence was measured using Raven's Standard Progressive Matrices (SPM), and SWL was measured using the SWL Scale. Age, gender, relationship status, political views, religion and the Facebook social network information were obtained from users' Facebook profiles. Users' consumption of alcohol, drugs and cigarettes and whether a user's parents stayed together until the user was 21 y old were recorded using online surveys. Visual inspection of profile pictures was used to assign ethnic origin to a randomly selected subsample of users. Sexual orientation was assigned using the Facebook profile "Interested in" field; users interested only in others of the same sex were labeled as homosexual, whereas those interested in users of the opposite gender were labeled as heterosexual.

Tecniche usate: Users and their Likes were represented as a sparse user–Like matrix binaria (1: associazione tra utente e Like; 0 altrimenti). The dimensionality of the user–Like matrix was reduced using singular-value decomposition (SVD). Numeric variables such as age or intelligence were predicted using a linear regression model, whereas dichotomous variables such as gender or sexual orientation were predicted using logistic regression. In both cases, we applied 10-fold cross-validation and used the k =100 top SVD components. For sexual orientation, parents' relationship status, and drug consumption only k = 30 top SVD components were used because of the smaller number of users for which this information was available. Vedi Allegato7.

Risultati ottenuti: Tratti dicotomici: alcuni predetti con grande accuratezza, altri più difficili. Sesso, Caucasici/Afroamericani precisi oltre al 90%. Vedi Allegato8. Tratti continui: predetti con grande accuratezza (usando il Pearson product-moment correlation coefficient) età, densità e dimensione della rete di amicizie su Facebook. SWL (l'umore può cambiare nel breve termine) e alcuni tratti di OCEAN non facilmente predicibili. I tratti psicologici possono solo essere misurati approssimativamente, a differenza di altri "numericamente esatti" (età, numero amici, ecc.). The transparent bars presented in Allegato9 indicate the accuracy of the questionnaires used as expressed by their test-retest reliabilities (Pearson product–moment correlation between the questionnaire scores obtained by the same respondent at two points in time).

I test sono stati ripetuti con un numero di Like sempre minore, ottenendo buone predizioni anche con pochi dati a disposizione. Vedi Allegato10.

Hanno infine visto quali sono le pagine Facebook più "discriminanti" per alcune feature ("Science -> High Intelligence; "Sephora" -> Low intelligence; "Shaq" (credo sia il cestista Shaquille O' Neil) -> Heterosexuality (confermo che si tratta del bestione di 2,16m ^_^); "Gay Marriage" -> Gay; ecc.)