Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea in Informatica

Verso l'identificazione automatica di utenti Twitter a rischio binge drinking

Relatore: Prof.ssa Gabriella Pasi

Correlatore: Dott. Marco Viviani

Relazione della prova finale di:
Luca Chiodini
Matricola 806976

Anno Accademico 2017-2018

A Davide

A Giuditta

ABSTRACT

Il fenomeno del "binge drinking", ovvero l'assunzione di diverse bevande alcoliche nel giro di poco tempo, è in preoccupante ascesa tra adolescenti e giovani. C'è quindi un interesse crescente, da parte di esperti e medici, nel tentare di prevenirlo e contrastarlo.

La comunità scientifica ha impiegato sinora metodi tradizionali per ricavare preziose informazioni sul problema, mentre resta inesplorata la possibilità di analizzare automaticamente e in modo intelligente la grande mole di dati generati dall'utilizzo massivo dei social media.

Questa relazione descrive l'attività di stage, il cui scopo è stato la definizione di metodi e algoritmi automatici per estrarre da Twitter informazioni sul binge drinking e identificare utenti potenzialmente a rischio, finalità particolarmente rilevante dal punto di vista della salute. Sono illustrati due approcci: il primo è finalizzato all'estrazione e all'analisi di dati, parole ed espressioni che più frequentemente compaiono nei tweet con un contenuto correlato all'alcol; il secondo approccio traccia un possibile percorso per arrivare alla classificazione degli utenti in potenziali "binge drinker" tramite classici algoritmi di apprendimento automatico.

(English version)

The phenomenon of “binge drinking”, which consists of intaking several alcoholic beverages in a short period of time, is growing at an alarming rate among teenagers. This has led to an increasing interest, by experts and doctors, in its prevention and contrast.

So far the scientific community has employed traditional methods to gather relevant information about the issue, while it is still unexplored the opportunity to automatically analyze the great volume of data generated by the widespread use of social media.

This report describes the work developed during the internship, whose aim was to define automatic methods and algorithms to extract from Twitter information about binge drinking and identify users having a potential risk, which is a particularly relevant purpose from the point of view of health. Two approaches are described: the former is intended to extract and analyze data, words and expressions appearing frequently in alcohol-related tweets are highlighted; the latter paves a possible way to classify users as potential “binge drinkers” using classical machine learning algorithms.

INDICE

1

INTRODUZIONE

1.1 IL FENOMENO DEL BINGE DRINKING

L'assunzione di alcol in quantità rilevanti costituisce da sempre un serio rischio per la salute. In tempi recenti, si è notato come l'abuso di questa sostanza da parte di giovani e giovanissimi sia in preoccupante crescita. Per definire un preciso tipo di comportamento, si è introdotto - nel mondo anglosassone prima, in Italia come prestito linguistico poi - il termine "binge drinking". La traduzione letterale dall'inglese può essere intesa genericamente come "bere in modo estremo"; esiste tuttavia un consenso pressoché unanime nel fissarne una definizione più rigorosa dal punto di vista medico: pratica binge drinking chi beve più di 5 unità alcoliche (dove un'unità alcolica, UA, è pari a 12 grammi di alcol puro) se uomo oppure più di 4 UA se donna [Wechsler et al. 1995].

L'ultimo rapporto ISTAT[1] del 2017 sul consumo di alcol in Italia riporta che più di 8 milioni di persone hanno avuto comportamenti che rientrano in questa definizione. La fascia di età 18-24, in particolare, è quella più a rischio: il 21,8% dei ragazzi e l'11,7% delle ragazze dichiarano di averlo praticato almeno una volta.

Da questi primi semplici dati è facile evincere come l'attenzione da parte di medici e psichiatri sia particolarmente alta in merito a questo fenomeno, per perseguire il duplice scopo di comprenderne anzitutto le caratteristiche e, in seconda istanza, di prevenirlo.

1 https://www.istat.it/it/files/2017/04/Consumo_alcol_in_Italia_2016.pdf

1.2 LA DIFFUSIONE DEI SOCIAL MEDIA E DEL CONTENUTO GENERATO DAGLI UTENTI

Con la diffusione delle tecnologie del Web 2.0, agli utenti è stata data la possibilità di passare dall'essere meri fruitori di contenuti a poterne essere autori. Questo processo di "democratizzazione" ha consentito un'ampia diffusione di ingenti quantità del cosiddetto contenuto generato dagli utenti, o UGC (User-Generated Content), attraverso molteplici piattaforme.

È sotto gli occhi di tutti come oggigiorno l'utilizzo dei social media sia diventato pervasivo in tutti gli ambiti e per tutte le fasce d'età della popolazione. Fare un elenco completo di tutti questi social network sarebbe impossibile; ricordiamo, tra i più famosi, Facebook, Instagram e Twitter. Quest'ultimo, in particolare, è un servizio che pone al centro dei brevi messaggi, noti in gergo come "tweet", che gli utenti pubblicano sui propri profili e attraverso i quali interagiscono tra loro. La quantità di dati generata è molto significativa ed essenzialmente di tipo non strutturato, essendo il cuore delle informazioni contenuto proprio nei tweet. Nel primo trimestre del 2018, negli Stati Uniti gli utenti attivi su Twitter sono stati circa 69 milioni [Rushe 2018]. A livello globale, stime recenti riferiscono di 400 milioni di tweet scambiati quotidianamente sulla piattaforma [Tsukayama 2013].

Diventa quindi fondamentale riuscire a gestire ed estrarre significato e conoscenza da una simile quantità di testo espresso in linguaggio naturale. Per quanto i linguaggi naturali abbiano una base di regole ben definite, determinate dalla morfologia e dalla sintassi di una lingua, il loro uso nella vita di tutti i giorni è continuamente soggetto a variazioni dovute a diverse cause (fattori sociali, culturali e di contesto, oltre all'evoluzione nel tempo). L'elaborazione del linguaggio naturale, nota anche come NLP (Natural Language Processing) [Chowdhury 2003], è la branca dell'informatica che si occupa della definizione di tecniche che permettono di analizzare varie dimensioni di un testo ed è attualmente considerata tra quelle che affrontano i problemi intrinsecamente più difficili ma anche, di riflesso, provano a risolvere le sfide più interessanti.

1.3 L'ANALISI DEL CONTENUTO PER SCOPI CLINICI: DALLE METODOLOGIE TRADIZIONALI A QUELLE DIGITALI

Diversi studi hanno dimostrato le potenzialità dell'elaborazione di grandi quantità di dati per analizzare diversi ambiti legati alla sfera della salute e della psichiatria, come lo smettere di fumare, i tentativi di suicidio e i disturbi dell'alimentazione. In queste ricerche sono stati utilizzati approcci detti "data-driven" - guidati dai dati - che possono essere un ottimo complemento a strategie più consolidate [Chen and Wojcik 2016].

Cercare di acquisire informazioni sul binge drinking non è certo una necessità nata solo ora. I metodi tradizionali prevedono in linea di massima la somministrazione di un questionario, la sua compilazione da parte di un ampio insieme di soggetti che costituiscono il campione, la catalogazione dei risultati e infine una fase di analisi attraverso metodi statistici.

L'équipe di psichiatria del dipartimento di Medicina e Chirurgia dell'Università di Milano-Bicocca ha effettuato in passato simili studi sui correlati del binge drinking [Bartoli et al. 2014] nei quali è stato somministrato un questionario tramite un'applicazione su smartphone a giovani adulti nella fascia 18-24 anni, incontrati all'esterno dei locali della vita notturna milanese.

I dati ottenibili con questa modalità sono interessanti ma comunque limitati: in particolare, le opinioni esprimibili sono limitate dalle scelte previste dal sondaggio e sono in un certo senso mascherate dalla coscienza stessa del soggetto che non si sente del tutto libero di fare alcune affermazioni che potrebbero metterlo in imbarazzo oppure sottoporlo a pressioni da parte della cerchia dei suoi pari.

Simili problematiche sono emerse in molti altri ambiti, tra i quali vi è l'attività di farmacovigilanza, in cui si cerca di catalogare in maniera sistematica tutti i potenziali effetti collaterali di farmaci in via di commercializzazione oppure già commercializzati. Anche in questo caso esiste un canale tradizionale costituito dai medici di famiglia, dai medici ospedalieri e dai farmacisti attraverso cui raccogliere informazioni; i pazienti spesso però dimenticano di comunicare tutti gli effetti che hanno riscontrato e sono

talvolta reticenti nel riportarli. Sono stati predisposti anche sistemi online anonimi per tentare di ovviare alle limitazioni esposte sopra, ma questi ultimi si sono dimostrati sottoutilizzati a causa della natura volontaria delle segnalazioni [O'Connor et al. 2014].

Proprio in merito alla farmacovigilanza, sono già state svolte alcune sperimentazioni [O'Connor et al. 2014] che si muovono nella direzione di utilizzare Twitter come sorgente di dati non filtrati: gli utenti scrivono spesso messaggi pubblici senza i freni inibitori descritti in precedenza.

Sulla base delle informazioni in nostro possesso, questo è il primo lavoro in cui l'elaborazione automatica di tweet viene applicata all'analisi del fenomeno del binge drinking. Il progetto trae le sue origini dagli studi svolti in precedenza dall'équipe di psichiatria dell'Ospedale San Gerardo di Monza, che afferisce all'Università di Milano - Bicocca, sul binge drinking [Bartoli et al. 2014; Carrà et al. 2015]. I risultati ottenuti sono stati possibili solo grazie alla collaborazione tra i due dipartimenti universitari, Medicina e Informatica.

1.4 ORGANIZZAZIONE DEL LAVORO

Nei capitoli successivi verranno illustrate le varie fasi dello studio svolto durante l'attività di stage, partendo dalla raccolta dei tweet per arrivare a una possibile strada per l'identificazione automatica degli utenti considerabili a rischio.

In particolare, nel Capitolo 2 sono descritte le modalità con cui sono stati scelti e scaricati i tweet e le ragioni per cui sono stati raggruppati in due dataset distinti. Nel Capitolo 3, dopo averne motivato la necessità, viene descritto un classificatore per separare gli utenti corrispondenti a persone fisiche dal resto dei profili. Nel Capitolo 4 sono illustrate e commentate alcune statistiche di alto livello a riguardo dei due dataset. Infine, nel Capitolo 5 viene descritta in dettaglio la progettazione di un secondo classificatore per individuare gli utenti a rischio binge drinking. Conclude questa relazione un'analisi critica dei risultati ottenuti.

2

APPROCCIO SVILUPPATO E CREAZIONE DEL DATASET

2.1 IL PROBLEMA CONSIDERATO

Come indicato nel capitolo introduttivo, l'attività di stage ha avuto come scopo l'analisi di dati provenienti da social media al fine di ricavare informazioni sul fenomeno del binge drinking. L'obiettivo è definire nuove procedure basate sui dati per svelarne significati non ricavabili manualmente, vista la quantità. Si presuppone infatti che vi sia una certa correlazione tra i contenuti pubblicati sui social network - Twitter nello specifico - e i comportamenti rischiosi per la salute, tra i quali vi è il binge drinking.

In [Dredze 2012] viene consigliato di approfondire la possibilità di utilizzare Twitter come mezzo per aumentare in modo significativo le potenzialità dei sistemi di sanità pubblica attuali. Seguendo quanto suggerisce l'articolo, in questo lavoro di stage si è partiti dallo scaricamento dei tweet per poterne effettuare successive analisi tramite appositi algoritmi.

2.2 CREAZIONE DEL DATASET

Il primo passo necessario per analizzare i tweet è naturalmente il loro reperimento. La scelta della modalità con cui selezionare i tweet è particolarmente rilevante per il seguito del lavoro, poiché tutte le attività successive saranno basate esclusivamente su questi testi.

L'équipe di psichiatria che ha collaborato a questo lavoro ha selezionato, prima dell'avvio di questa tesi, alcuni hashtag che si ritiene vengano utilizzati da utenti che

potrebbero inviare messaggi relativi al fenomeno del binge drinking. La ricerca dei tweet è limitata a quelli scritti in lingua inglese (così come catalogati da Twitter stesso). Tramite ricerche manuali su Twitter, i risultati della ricerca attraverso ciascuno di questi hashtag sono stati brevemente vagliati al fine di valutarne l'opportunità di inclusione nel dataset finale. È stata a tal fine condotta una breve ricerca pilota per sette giorni osservando i tweet raccolti.

Sono stati individuati 23 hashtag, che ricordiamo essere le "parole chiave" spesso aggiunte nei tweet per facilitarne la ricerca, con diversi gradi di dettaglio:

- hashtag che riguardano bevande alcoliche in forma generica: `#cocktail`, `#alcohol`, `#vodka`, `#rum`, `#drinks`;
- hashtag che indicano fenomeni noti per essere spesso forme di abuso di alcol, ad esempio "pub crawl" che indica l'azione del bere in diversi pub nella stessa sera: `#pubcrawl`, `#pubcrawling`, `#botellon`;
- hashtag che indicano in modo esplicito postumi comuni dell'aver bevuto troppo alcol oppure la situazione stessa di ubriachezza: `#wasted`, `#hangover`, `#toomuchalcohol`, `#sorehead`, `#drunkies`, `#drunkasfuck`;
- hashtag che contengono un riferimento diretto al binge drinking: `#bingedrinking`.

L'elenco completo dei 23 hashtag è il seguente:

`#bingedrinking`	`#alcohol`	`#alcoholic`	`#alcoholics`
`#nomorealcohol`	`#cocktail`	`#cocktails`	`#drinks`
`#drinking`	`#vodka`	`#rum`	`#wasted`
`#hangover`	`#toomuchalcohol`	`#drunk`	`#drunkasfuck`
`#drunkennights`	`#drunkies`	`#getdrunk`	`#sorehead`
`#pubcrawl`	`#pubcrawling`	`#botellon`	

Un primo crawler, sviluppato da un membro del laboratorio di Information Retrieval del dipartimento di Informatica, è stato inizialmente utilizzato per scaricare da Twitter, tramite API pubbliche, tutti i tweet in lingua inglese che contenessero uno dei 23

hashtag definiti; contestualmente sono state scaricate anche le informazioni di base dal profilo dei rispettivi autori. Tale processo ha raccolto 452 262 tweet appartenenti a 160 557 utenti unici, in un periodo compreso tra il 29/11/2017 e il 22/03/2018.

Durante una successiva ispezione del file CSV contenente questo dataset iniziale si è notato come la maggioranza dei tweet fosse monca della parte finale. Da un breve approfondimento è emerso che a partire da novembre 2017, dopo alcuni mesi di test, Twitter ha modificato la dimensione massima prevista per ciascun messaggio, portandola da 140 a 280 caratteri. Questa modifica è intercorsa durante lo sviluppo del primo crawler, che era stato quindi progettato per scaricare tweet di 140 caratteri al massimo, provocando dunque lo scaricamento solo parziale del testo dei tweet. In questo lavoro si è quindi sviluppato un piccolo script ad-hoc in Python che, sfruttando la libreria tweepy[1], riscaricasse l'intero testo sfruttando le nuove API con risposte in formato JSON.

Affinché lo scaricamento dei tweet non si interrompesse e per poter fare in futuro confronti tra due dataset appartenenti a periodi diversi, si è anche sviluppato da zero un nuovo crawler che utilizza anch'esso la libreria tweepy per scaricare ogni ora nuovi dati da Twitter. Questo secondo crawler è stato avviato il 16 aprile 2018 e ha raccolto (dato aggiornato al termine del mese di maggio) ulteriori 194 815 tweet.

Nel dettaglio, i dati scaricati e conservati per ciascun tweet sono:

- il testo completo, anche nel caso in cui superi i 140 caratteri originali oppure sia un retweet (recuperando il testo iniziale, poiché quello retweetato viene troncato);
- metadati associati al tweet: numero di reazioni (numero di retweet e di like), data e ora di creazione, geolocalizzazione laddove presente;
- qualora il tweet sia una risposta ad un altro messaggio oppure qualora sia un retweet (ovvero, la semplice riproposizione del tweet di un altro utente), vengono conservate informazioni sul tweet originale e sul nome dell'autore di quest'ultimo;

1 http://www.tweepy.org

- metadati relativi all'autore del tweet: screenname (detto anche handle, ovvero il nickname scelto dall'utente), nome completo, testo della biografia nel suo profilo, numero di tweet pubblicati nella storia, numero di following e followers, presenza di un URL nell'apposito campo della biografia, data di creazione del profilo.

Per dare un'idea dei dati scaricati, riportiamo di seguito due tweet esemplificativi estratti dal dataset. Sebbene il formato scelto per la memorizzazione sia un file CSV per ragioni di semplicità e retrocompatibilità, gli esempi sono mostrati in una notazione JSON interna per aumentarne la leggibilità.

```
{
  screen_name: 'PasiniDebora',
  hashtag: '#alchohol',
  text: 'Read. Eat. Drink. R E D. #feltrinelli #feltrinellired #spritz
      #aperolspritz #red #read #eat #drink #alcohol #friends #weekend
      #saturday #milan #milano #citylife... https://t.co/s4eHxXuhDn',
  favorite_count: 0,
  in_reply_to_screen_name: '',
  created_at: 1530365362,
  in_reply_to_status_id: '',
  tweet_id: '1013082069860220929',
  in_reply_to_user_id: '',
  retweet_count: 0,
  retweeted: False,
  geolocation_lon: 9.15702,
  geolocation_lat: 45.47827,
  user_description: 'Struggling on blindly, never getting anywhere and
      hoping to die before I realise I never made it | Wannabe writer,
      geek extraordinaire, everyday shipper',
  user_statuses_count: 10965,
  user_followers_count: 226,
  user_favorites_count: 29028,
  user_following_count: 465,
  user_description_url: 'https://t.co/vVYwRhRNn1',
  user_name: 'Debora Pasini',
  user_created_at: 1368977974,
  user_location: 'Lombardia, Italia',
  lang: 'en',
  user_id: '1441789952'
}
```

Codice 1: Esempio dei dettagli memorizzati per un tweet geolocalizzato, pubblicato da un utente che indica nel proprio profilo URL e posizione. I campi `created_at` e `user_created_at` sono in formato Unix timestamp.

```
{
  screen_name: 'Custardboy2',
  hashtag: '#alchohol',
  text: '@cigarboyrick82 I hate to say it, but he will probably do it
      again #addiction #alcohol',
  favorite_count: 0,
  in_reply_to_screen_name: 'cigarboyrick82',
  created_at: 1523889258,
  in_reply_to_status_id: '985887071247183872',
  tweet_id: '985919320504111104',
  in_reply_to_user_id: '2226783061',
  retweet_count: 0,
  retweeted: False,
  geolocation_lon: '',
  geolocation_lat: '',
  user_description: 'foodie,football,family - remember,banter is a good
      thing,all thoughts are my own and probably wrong',
  user_statuses_count: 13391,
  user_followers_count: 433,
  user_favorites_count: 11277,
  user_following_count: 1103,
  user_description_url: '',
  user_name: 'Simon Richards',
  user_created_at: 1439393563,
  user_location: '',
  lang: 'en-GB',
  user_id: '3418342065'
}
```

Codice 2: Esempio dei dettagli memorizzati per un tweet scritto in risposta a un altro utente, con i riferimenti al tweet e all'utente originali.

3

STATISTICHE E CLASSIFICAZIONE DEGLI UTENTI FISICI

3.1 STATISTICHE GENERALI

Come illustrato nel capitolo precedente, siamo ora in possesso di due dataset distinti composti da alcune decine di migliaia di tweet con un contenuto ragionevolmente correlato all'alcol e al binge drinking.

Può rivelarsi interessante analizzare subito alcune caratteristiche salienti di questi due gruppi di tweet. Nel prosieguo di questo documento, ci riferiremo per comodità al primo dataset con D1 e al secondo con D2.

	D1	D2
Data inizio	29/11/2017	16/04/2018
Data fine	22/03/2018	31/05/2018
Giorni	114	46
Numero di tweet	452 262	194 815
Numero medio di tweet giornalieri	3 967	4 235
Autori unici	160 557	88 987
Numero medio di tweet per autore	2,81	2,19

Tabella 1: Statistiche di base sui due dataset.

Queste statistiche basilari mostrano già due fatti interessanti rispetto agli obiettivi iniziali del progetto:

- Il numero medio di tweet giornalieri è lievemente maggiore in D2 (+6,8%) rispetto a D1. Il numero di tweet subisce quindi l'effetto della stagionalità: il primo dataset è essenzialmente relativo al trimestre invernale, mentre il secondo

al periodo di fine primavera. Si può quindi ipotizzare che l'approssimarsi della stagione estiva sia correlato con un aumento del consumo di bevande alcoliche.

- Il numero medio di tweet per autore è drasticamente minore in D2 (-22%) rispetto a D1. Questa forte osservazione, combinata con quella del punto precedente, corrobora l'ipotesi dell'aumento del numero di utenti che hanno pubblicato tweet relativi all'alcol nel periodo primaverile.

Unendo le due osservazioni, concludiamo che nel periodo primaverile ciascun utente ha pubblicato (in media) un numero minore di tweet; ciononostante, il numero complessivo di messaggi è cresciuto. Uno studio ulteriore relativo a un periodo estivo potrebbe avvalorare maggiormente l'ipotesi del consumo maggiore di bevande alcoliche nella stagione estiva.

3.2 CHI SONO GLI AUTORI DEI TWEET?

Prima di proseguire, è fondamentale interrogarsi un po' più a fondo sulla natura dei dati che abbiamo raccolto nei due dataset. Se è vero che Twitter può rivelarsi una preziosa fonte di contenuti pubblicati da parte di utenti reali, ovvero persone fisiche che utilizzano la piattaforma per diffondere in maniera trasparente i propri commenti, interessi, opinioni e comportamenti; d'altra parte non è irrilevante l'attività di spammer, bot e attività commerciali che sfruttano lo stesso social network con scopi ben differenti. Le finalità per cui questi tweet vengono pubblicati sono molto diverse tra loro: alcuni tweet includono messaggi positivi, come offerte di aiuto in caso di problemi legati all'alcol oppure video a carattere educativo su YouTube; altri neutri, come notizie di stampo giornalistico o blog con contenuti legati all'alimentazione; altri ancora hanno una connotazione negativa rispetto al problema del binge drinking, ad esempio account Ebay che promuovono la vendita in stock di alcolici oppure baristi che invogliano a provare nuovi cocktail.

Risulta quindi necessario anzitutto comprendere e poi rimuovere questa "fonte di rumore", pena il raggiungimento di risultati grossolani [Chen and Wojcik 2016].

Consideriamo a titolo esemplificativo alcuni tweet. I due seguenti hanno come finalità la prevenzione del consumo di alcol tra gli adolescenti: il primo proviene da uno sportello online, mentre il secondo da un esponente di un'associazione australiana che studia la diffusione delle bevande alcoliche tra adolescenti.

> Quitting #drinking won't be easy, but you can do it!
>
> — *@AlcholDisease*

> Thanks for a great welcome #ANU. Students were keen yesterday to learn about our campaign. We'll be around for the rest of the year to help reduce the harm to students from #alcohol. @anustudents @ANUmedia @RichardmBaker @Woroni @ourANU @ANUsport @ACTHealth @FAREAustralia
>
> — *@MichaelTThorn*

Ci sono attività commerciali (bar, pub, discoteche, hotel) che si fanno pubblicità con profili "aziendali" oppure tramite i loro barman, spesso obbligati a mettere esplicita indicazione sul divieto di consumo da parte di minorenni.

> $5 #appetizers, $2 Domestic #beer bottles and well #drinks : 4p.m. - 6p.m. #happyhour #BOOM
>
> — *@CityTapTheAttic*

> Crown Royal, lemonade, and honey! What an awesome warm weather cocktail! #diageorep #crownroyal #cocktails #drinkresponsibly
>
> — *@bottleswbrynne*

Infine, ci sono tweet che citano l'alcol per i più svariati motivi. Il seguente, per esempio, proviene da un account di una testata giornalistica.

> LIVE UPDATE: Second #school employee arrested for distributing #alcohol to Battery Creek H - Apr 16 @ 3:22 PM ET
>
> — *@PulpNews*

Una prima fase dell'analisi dei dati si è quindi concentrata sullo scartare tutti quei profili che non rientrano nella definizione di "persona fisica che utilizza l'account senza finalità commerciali".

3.3 CLASSIFICAZIONE DEGLI UTENTI

Esistono alcuni approcci descritti in letteratura per affrontare il problema di distinguere utenti reali da bot e simili [Chu et al. 2012; Igawa et al. 2016; Guo and Chen 2014]. Twitter stesso è in lotta da tempo contro un'attività impressionante in termini di tweet prodotti da parte di spammer e account falsi [Thomas et al. 2011]. Il compito si riduce essenzialmente a una classificazione binaria degli account in "reali" e non. Naturalmente, la parte cruciale è la scelta delle caratteristiche, vale a dire delle feature, associate agli account considerati e al contenuto dei loro tweet. In riferimento al problema considerato, ne sono state individuate undici, di cui diamo singolarmente una ragione dettagliata per la quale è stata considerata.

3.3.1 *Descrizione delle feature*

- **Numero di tweet associati a un profilo**. Gli utenti con un alto numero di tweet sono probabilmente attività commmerciali oppure bot [Chu et al. 2012].
- **Numero medio di hashtag per tweet**. Gli hashtag sono le "parole chiave" utilizzate dagli utenti per categorizzare il loro messaggio. Ci si aspetta che un utente includa un numero limitato di hashtag in un singolo tweet, mentre chi vuole promuovere il proprio contenuto spesso ne fa un uso sovrabbondante, abusan-

done, per comparire in tutte le possibili ricerche tramite quegli hashtag [Bian et al. 2012; Guo and Chen 2014].

- **Numero medio di menzioni per tweet**. Le menzoni, ovvero l'utilizzo della '@' seguita dal nome di un altro utente, sono utilizzate per la conversazione e la discussione su Twitter. Queste interazioni sono maggiormente proprie delle persone reali, mentre le attività commerciali lanciano spesso messaggi diretti a tutto il pubblico e non intrattengono conversazioni individuali con la propria cerchia di seguaci [Bian et al. 2012].

- **Numero medio di pronomi personali per tweet**, il cui utilizzo è strettamente connesso alle persone: i messaggi pubblicitari sono scritti spesso in forma asciutta e impersonale [Bian et al. 2012].

- **Numero medio di URL citati per tweet**. Collegamenti ipertestuali che rimandano a siti esterni (spesso più di uno) sono frequentemente inseriti dalle attività commerciali per spostare la navigazione degli utenti da Twitter a quella del loro brand. I risultati sperimentali in [Chu et al. 2012] mostrano che questa è la seconda feature maggiormente discriminante tra bot e utenti reali.

- **Rapporto tra il numero di retweet e il numero totale di tweet catalogati**. Un utente reale difficilmente include continuamente tweet altrui, senza commento, nel proprio profilo. Questo comportamento è invece tipico di alcuni account che retweetano apprezzamenti su un marchio oppure di agenzie giornalistiche o, ancora, di account che si comportano in stile feed RSS [Guo and Chen 2014].

- **Presenza dell'URL nei dettagli del profilo**. Nella pressoché totalità dei casi le attività commerciali sfruttano a pieno le potenzialità della piattaforma, indicando il proprio sito internet nell'apposito campo.

- **Dimensione della rete sociale**, intesa come somma del numero di follower e followee. Profili con almeno uno tra numero di followee e follower molto alto sono evidenza di una persona famosa oppure di un'azienda.

- **Rapporto tra numero di follower e followee**. Per gli account di utenti reali questo rapporto non si discosta troppo dall'unità: è ragionevole aspettarsi che una persona segua un certo numero di profili e che sia ricambiata da altrettanti. Spes-

so lo squilibrio è forte per persone famose e attività commerciali, che tendono ad avere un elevato numero di follower (anche nell'ordine delle decine o centinaia di migliaia di unità) ma pochissimi o addirittura zero followee (perché lo scopo di quell'account non è leggere i contenuti pubblicati da terzi). Una metrica molto simile a questa è adottata in [Chu et al. 2012] e traccia una linea di demarcazione significativa tra le due classi.

- **Presenza di almeno un tweet geolocalizzato** (in [Guo and Chen 2014] si parla dettagliatamente di classificazione tramite geolocalizzazione). L'utilizzo di Twitter avviene prevalentemente tramite app da smartphone, i quali hanno spesso la localizzazione attiva; d'altro canto la fruizione da desktop è una prerogativa in buona parte dell'utenza business. Consideriamo quindi con attenzione il fatto che di un utente disponiamo anche solo di un tweet geolocalizzato.
- **Modello bag-of-words**. In aggiunta alle feature descritte sopra, sono state individuate tramite ispezione manuale di un campione casuale di utenti alcune parole che indicano con un'alta probabilità un profilo non personale. Tali parole vengono ricercate all'interno del testo dei tweet (Tabella 2), nella descrizione dell'utente (Tabella 3) e come sottostringhe del suo nickname (Tabella 4).

3.3.2 *Calcolo delle feature*

Le feature sono calcolate in modo naïve tramite uno script Python, il cui pseudocodice è riportato nell'Algoritmo 1. Merita un commento specifico la divisione del testo in parole, perché l'operazione può sembrare più banale di quanto lo sia in realtà. Lo script fa uso del framework per l'elaborazione del linguaggio naturale in Python NLTK (Natural Language Toolkit)[1].

Un testo T può essere visto come una sequenza $t_1, \ldots, t_N$ di token. Definiamo token

1 https://www.nltk.org

sponsor	freetickets	dutyfree	free	sponsored
recipe	addiction	gift	treatment	shop
abuse	disease	stop	win	recovery
book	masterclass	quit	sobriety	quitting
addict	tutorial	shipped	t-shirt	hotline
discount	motivation	ad	illness	official
page	magazine	business	marketing	bar

Tabella 2: Elenco di token da cercare all'interno del testo dei tweet.

travel	recipes	organisation	shipped	recovery
events	help	shop	gifts	discounts
fitness	crowdfunding	treatment	free	charity
advertising	reservations	boutique	dependence	addiction
follow	distillery	official	page	magazine
business	marketing	bar	commercial	game
store	blog	inquires	promotional	corporate
editor	gin	prosecco	wodka	

Tabella 3: Elenco di token da cercare all'interno della descrizione di un utente.

journal	blog	hotel	disease	recovery
shop	magazine	fitness	natural	grocery
bar	recipe	spotlight	performance	marketing
addiction	news	renascence	book	travel
food	meal	social	tweet	distillery
country	town	lifestyle	official	magazine
win	game	bot	drink	business

Tabella 4: Elenco delle sottostringhe da cercare all'interno del nickname (`screen_name`) di un utente.

Algoritmo 1 Calcolo delle feature.

```
procedure ComputeFeatures(users)
    for all user ∈ users do
        ComputeUserFeatures(user)

procedure ComputeUserFeatures(user)
    hashtags, mentions, pronouns, url, retweet, bad_tokens ← 0
    geolocated_tweet ← False
    for all token ∈ tokenize(user.description) do                ▷ NLTK Tokenizer
        if token ∈ description_bad_tokens then
            bad_tokens ← bad_tokens + 1
    for all bad_screenname_string ∈ bad_screenname_strings do
        if bad_screenname_string.is_substring(user.screenname) then
            bad_tokens ← bad_tokens + 1
    for all tweet ∈ user.tweets do
        if tweet.is_retweet then
            retweet ← retweet + 1
        if tweet.has_geolocation then
            geolocated_tweet = True
        tokens ← tokenize(tweet)                                 ▷ NLTK Tokenizer
        pos_tagger(tokens)                                       ▷ NLTK Part-Of-Speech tagger
        for all token ∈ tokens do
            if token.part_of_speech = "PRP" then
                pronouns ← pronouns + 1
            if token.startsWith('#') then
                hashtags ← hashtags + 1
            else if token.startsWith('@') then
                mentions ← mentions + 1
            else if token ∈ tweet_bad_tokens then
                bad_tokens ← bad_tokens + 1
            else if token.is_URL then                            ▷ Regex
                url ← url + 1
    n ← length(user.tweets)
    hashtags_avg ← hashtags/n
    mentions_avg ← mentions/n
    pronouns_avg ← pronouns/n
    url_avg ← url/n
    retweet_ratio ← retweet/n
    statuses ← n
    network_size ← user.followers + user.following
    followers_ratio ← user.followers/(user.following + 0.001)
    return statuses, bad_tokens, hashtags_avg, mentions_avg,
        ↪ pronouns_avg, url_avg, retweet_ratio, user.has_description_url,
        ↪ network_size, geolocated_tweet, followers_ratio
```

in modo impreciso come un termine o una parola indivisibile. In alcuni casi il processo di divisione in token è banale:

$$\text{To} \mid \text{be} \mid \text{or} \mid \text{not} \mid \text{to} \mid \text{be}$$

ma in altri lo è molto meno: consideriamo *aren't* come esempio triviale. Deve essere separato (verbo più negazione) oppure no?

Fortunatamente NLTK include un tokenizer[2] (separatore in token), utilizzato in questo progetto per il modello bag-of-words, e un POS (part-of-speech) tagger[3], ovvero un modulo che processa i token e assegna probabilisticamente a ognuno la parte del discorso che svolge. Per computare il numero medio di pronomi personali per tweet è quindi sufficiente contare il numero di token etichettati come `PRP` (personal pronoun) e dividere per la quantità di tweet a disposizione per un certo profilo.

3.3.3 *Apprendimento automatico supervisionato*

I valori delle feature calcolati come indicato nella Sottosezione 3.3.2 sono ovviamente molto diversi tra i vari utenti. È quindi palesemente impossibile stabilire a priori dei valori limite per classificare i profili, in quanto ciascuno ha le proprie peculiarità. Si consideri come esempio lampante la dimensione della rete sociale: nessuna soglia potrebbe mai discriminare correttamente tutti gli account, visto l'elevato grado di variabilità. Per classificare gli utenti si è pertanto deciso di utilizzare tecniche di apprendimento automatico, applicate alle feature descritte in precedenza. In particolare, impiegheremo tecniche bastate su apprendimento supervisionato, per le quali è essenziale avere a disposizione un dataset etichettato, che costituisce la cosiddetta groundtruth, con cui è possibile costruire il modello e valutare la bontà dell'approccio.

Per questo progetto sono stati etichettati a mano dagli esperti dell'équipe di psichiatria dell'Ospedale San Gerardo di Monza $N = 494$ utenti, analizzandone il profilo. In caso di dubbio sull'attribuzione dell'etichetta, si è presa di volta in volta una decisione a

2 https://www.nltk.org/api/nltk.tokenize.html
3 https://www.nltk.org/api/nltk.tag.html

maggioranza. A ciascun profilo è stata apposta una duplice etichetta. La prima distinzione è stata fatta tra "utenti reali" e non: l'etichetta '1', corrispondente ai primi, è stata attribuita a $N' = 180$ utenti; l'etichetta '0', che comprende invece tutti gli altri casi, è stata attribuita ai rimanenti 314 utenti. In seconda istanza, restringendo l'attenzione ai primi $N' = 180$ utenti, si è verificata la possibilità che fossero a rischio di binge drinking, apponendo una seconda etichetta con l'esito di questa valutazione ('1' per indicare un soggetto potenzialmente a rischio, '0' viceversa). Questo secondo insieme di dati annotati verrà utilizzato nel Capitolo 5. In generale, tutte queste informazioni costituiscono il training set.

Vista la limitatezza del campione in esame, per valutare i risultati del classificatore si è scelto di utilizzare la tecnica della convalida incrociata (cross-validation), nello specifico una k-fold cross validation, con un valore di k uguale a 5. Con questa strategia, il training set viene diviso in cinque parti. Il classificatore viene valutato per cinque volte, utilizzando una parte come validation set su cui misurare i risultati e le rimanenti quattro come training set. Al termine, viene considerata media e deviazione standard per i vari parametri tra le cinque esecuzioni.

Per tutte le elaborazioni di apprendimento automatico è stato utilizzato il framework scikit-learn[4] (abbreviato in `sklearn`) che è divenuto lo standard de-facto per applicazioni di machine learning in Python.

Vista la diversità di scala dei valori associati alle feature, prima di essere analizzati da un classificatore i dati sono stati normalizzati in un intervallo comune tramite i metodi della classe `preprocessing.StandardScaler` del framework `sklearn`. In particolare, le funzioni obiettivo dei vari classificatori assumono implicitamente, per funzionare in modo corretto, che i dati siano centrati attorno allo zero e che abbiano uguale varianza. Per queste ragioni tale classe normalizza le feature affinché assomiglino a una distribuzione normale standard con media nulla e varianza unitaria. Formalmente, per ciascuna feature ogni valore x_i viene trasformato applicando la seguente formula:

$$x_i' = \frac{x_i - \mu_x}{\sigma_x}$$

4 http://scikit-learn.org

dove μ_x e σ_x sono rispettivamente media e deviazione standard dei dati di una certa feature.

Sono stati utilizzati e valutati due diversi classificatori basati su apprendimento supervisionato, seguendo approcci consolidati in letteratura [Chu et al. 2012; Bian et al. 2012; Guo and Chen 2014]:

- Macchine a vettori di supporto, o SVMs (Support Vector Machines), in cui si cerca di risolvere un problema di ottimizzazione: trovare un iperpiano che separi con il massimo margine le feature nel loro spazio dimensionale. Inoltre, viene utilizzata una funzione kernel che mappa i dati in uno spazio con più dimensioni al fine di trovare il migliore iperpiano di separazione. Il mapping può essere di tipo non lineare (ad esempio con un kernel RBF, di tipo gaussiano).

- Foresta casuale, o Random forest. È un classificatore di tipo ensemble, cioè appartenente alla categoria che utilizza più algoritmi di machine learning di base per ottenere predizioni migliori di quelle ottenibili dai singoli algoritmi presi individualmente. Nella fattispecie, random forest utilizza come modelli di base gli alberi di decisione.

  Un classificatore ad albero di decisione (decision tree), come dice il nome stesso, costruisce un albero alle cui foglie sono presenti le etichette delle classi e alle cui diramazioni viene intrapreso un certo ramo a seconda del valore di una o più feature. Questo modello ha il pregio di essere semplice ma tende ad adattarsi eccessivamente ai dati osservati nella fase di training (comportamento detto overfitting). In altre parole, un albero di decisione tende spesso a costruire regole per le diramazioni troppo specifiche, che non catturano proprietà generali e che quindi sono poco adatte per classificare accuratamente nuovi dati.

  Random forest prevede, durante la fase di training, la costruzione di un certo numero di alberi di decisione, ciascuno addestrato su un sottoinsieme causale di feature. L'attribuzione dell'etichetta finale è effettuata sulla base delle etichette prodotte da ogni albero indipendentemente, prendendo quella che occorre il maggior numero di volte. Così facendo il classificatore foresta casuale mitiga significativamente il problema dell'overfitting.

Poiché c'è un lieve squilibrio nella numerosità del campione tra le due classi, utilizziamo in modo implicito la tecnica standard del sovracampionamento (oversampling) della classe più piccola fino a raggiungere la parità (tramite il parametro `class_weight`).

I risultati dei due classificatori in termini di accuracy, precision, recall e F1 score sono riportati nella Tabella 5.

	Accuracy	Precision	Recall	F1 score
SVMs	0.76	0.68	0.65	0.66
Random Forest	0.73	0.68	0.50	0.57

Tabella 5: Confronto tra i risultati dei classificatori.

Nelle figure 1 e 2 viene mostrata la curva ROC (Receiver Operating Characteristic) ottenuta mediante l'utilizzo del primo e del secondo classificatore.

Applicando il classificatore SVMs (essendo quello che ottiene i risultati migliori) agli interi dataset D1 e D2 otteniamo due sottoinsiemi di utenti etichettati come "persone fisiche".

	Training set	D1	D2
Utenti totali	494	160 557	88 987
Utenti etichettati come "reali"	180	70 290	40 181
% su totale	36,4 %	43,8 %	45,2 %

Tabella 6: Esito della classificazione con SVMs su D1 e D2.

I risultati ottenuti in termini di accuracy (cfr. Tabella 5) sono discreti. In [Chu et al. 2012] la classificazione raggiunge livelli che superano il 90% di accuratezza, ma il compito svolto è intrinsecamente più semplice. Nel paper viene infatti mostrato un approccio per distinguere bot e spammer, che costituiscono solo un sottoinsieme di quanto identificato in questo lavoro, che ha l'obiettivo più ambizioso di riconoscere e separare anche attività commerciali e profili utilizzati per finalità pubblicitarie.

I risultati sono comparabili con quelli ottenuti in [Bian et al. 2012], in cui si identificavano con accuracy = 0,74 gli utenti Twitter che stavano assumendo un qualche tipo di farmaco.

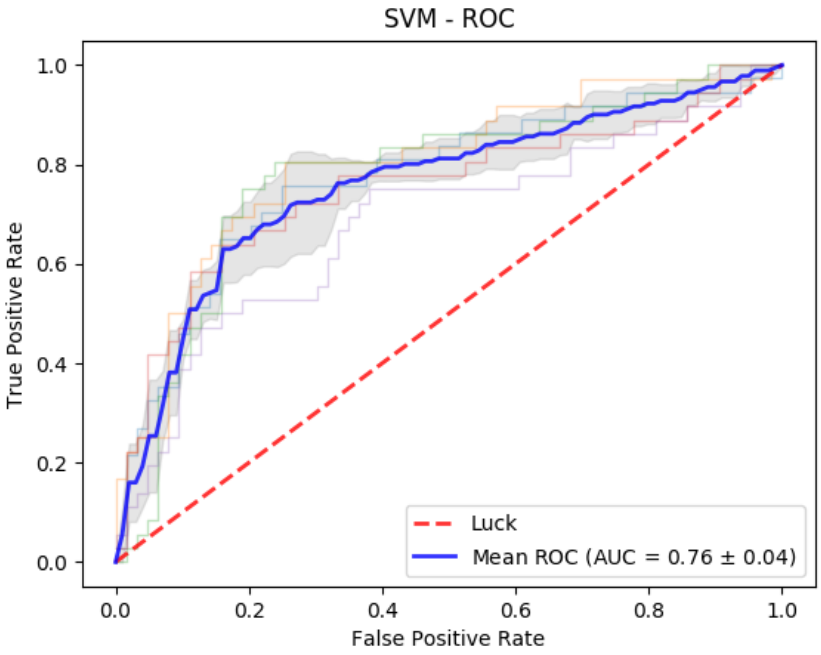

Figura 1: Valore medio AUC e curva ROC del classificatore SVM.

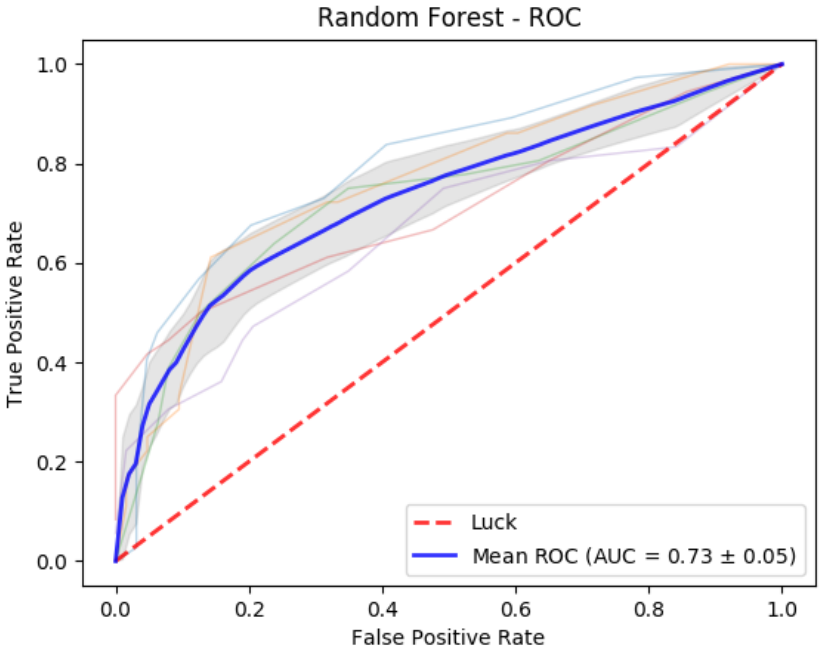

Figura 2: Valore medio AUC e curva ROC del classificatore Random Forest.

4

DI COSA SI PARLA NEI TWEET?

La fase di classificazione, illustrata nel capitolo precedente, è servita a ripulire il dataset dalla maggior parte dei tweet che esulano dall'interesse specifico di questa ricerca, ovvero quelli pubblicati da attività commerciali, account di stampo giornalistico, pubblicitari e così via.

Può essere interessante comprendere a questo punto "di cosa si parla" su Twitter quando i messaggi sono relativi all'alcol. Per far questo, separatamente per ciascun dataset, sono stati estratti i 50 n-grammi più frequenti.

Data una sequenza ordinata di token $T = t_1, \ldots, t_k$, è detto n-gramma (con $n \geqslant 1$) ciascuna sottosequenza di token contigui $t_a, \ldots, t_b$ lunga esattamente n. In questo lavoro ci siamo limitati a considerare unigrammi ($N = 1$), bigrammi ($N = 2$) e trigrammi ($N = 3$).

Nella nuvola di parole mostrata in Figura 3 sono riportati i cinquanta unigrammi più frequenti nel dataset D2, escluse le stopwords (ovvero le parole che appaiono più comunemente in quanto elementi costituitivi delle frasi, come 'the' o 'and').

Non si registrano differenze significative tra i due dataset, i cui n-grammi più frequenti sono essenzialmente gli stessi. È comunque interessante osservare quali sono i bigrammi e i trigrammi più frequenti (cfr. Tabella 7).

È immediatamente evidente la presenza di numerose parole aventi un'accezione generalmente positiva, come "live", "life" e "love". Ci sono alcune espressioni che provengono con ogni probabilità da profili di persone che si occupano a vario titolo (giornalisti, medici, esperti) di aiutare persone in difficoltà: "mental health", "public health", "need heelp" e "help tweet". Quest'ultima indica nello specifico che Twitter viene già sfruttato come punto di primo contatto tra chi offre aiuto a riguardo dell'alcol e chi

Figura 3: I 50 unigrammi più frequenti in D2, stopwords escluse.
La dimensione dell'unigramma è proporzionale alla sua frequenza.

Bigramma	# occorrenze
i love	648
fan account	242
i like	213
all views	184
live life	141
social media	126
mental health	124
animal lover	120
public health	118
love life	106
need help	98
help tweet	84
food drink	66
food wine	65
i enjoy	62

Trigramma	# occorrenze
need help tweet	94
help tweet us	94
must 21 follow	47
share anyone 21	36
work hard play	26
please drink responsibly	25
i love music	20
all views expressed	20
i love travel	17
i love family	16

Tabella 7: Bigrammi e trigrammi più frequenti in D2.
Nota: sono stati inclusi solo elementi ritenuti significativi dopo un vaglio manuale.

ne ha bisogno. Infine, colpiscono l'attenzione i bigrammi "food drink" e "food wine": si ritiene che possano provenire da persone che stanno effettivamente bevendo alcolici.

Osservando i trigrammi si conferma la stessa tendenza in merito all'aiuto di persone in difficoltà ("need help tweet") e si aggiunge un fenomeno singolare: diversi profili che pubblicano spesso tweet in materia di alcolici, ad esempio ragazzi che lavorano nei fine settimana come barman, includono nei loro tweet un'indicazione esplicita al divieto per i minori di consumare alcol (l'età legale minima in America è di 21 anni); ne sono testimonianza le espressioni "(don't) share anyone 21", "must (be) 21 (to) follow" e "please drink responsibly".

5

INDIVIDUAZIONE DEI POTENZIALI BINGE DRINKER

In questo capitolo viene descritta la fase cruciale che ha motivato l'intero progetto, nonché la più complessa: l'individuazione degli account appartenenti a persone che potrebbero praticare o aver praticato binge drinking.

È doverosa una premessa delicata ma fondamentale: il binge drinking è riconosciuto essere uno tra i più grandi problemi di sanità pubblica; la sua diagnosi è pertanto di competenza di personale medico psichiatrico, a cui spetta la valutazione finale tenuto conto di una serie di fattori. Quel che si vuole portare avanti in questo progetto è un'attività di supporto al lavoro di questi professionisti, tramite le possibilità offerte dai più recenti progressi in ambito di machine learning e text mining, analizzando il contenuto di una serie di tweet.

L'assunzione di fondo per l'individuazione dei soggetti interessati dal fenomeno è che essi condividano almeno in parte lessico ed espressioni utilizzate nei tweet. Questa convinzione può sembrare un po' forte alla luce del "rumore di fondo" presente su Twitter: non ci si aspetta di certo che un utente pubblichi continuamente messaggi a riguardo di questo tema. In aggiunta, le espressioni gergali e le limitazioni sul numero dei caratteri massimi di un tweet hanno un forte effetto sul modo in cui vengono scritti i tweet, rendendone complessa l'analisi.

Date queste premesse, si è sviluppato in questa fase un nuovo classificatore che si concentra sul contenuto dei tweet, tenendo in considerazione feature esclusivamente linguistiche. Già in [Bian et al. 2012] si osservava come fosse necessario approfondire l'analisi di feature testuali e semantiche ai fini di una classificazione efficace.

5.1 TF-IDF

In questa sezione forniamo alcune definizioni utili a costruire feature per questo nuovo classificatore.

Dato un documento d e un termine t, chiamiamo *term frequency* $\mathrm{tf}(t, d)$ il numero di occorrenze di t in d. Il valore di tf sarà tanto maggiore quante più volte un certo termine appare in un documento. Analogamente definiamo *document frequency* $\mathrm{df}(t)$ il numero di documenti che contengono il termine t.

Data una collezione di N documenti, chiamiamo *inverse document frequency* di un termine t il valore

$$\mathrm{idf}(t) = \log \frac{N}{\mathrm{df}(t)}$$

Una volta calcolati tf e idf, è possibile definire la funzione di peso $\mathrm{tf\text{-}idf}$ (*term frequency - inverse document frequency*) che permettere di misurare l'importanza di un termine in un documento rispetto all'intera collezione di documenti:

$$\mathrm{tf\text{-}idf}(t, d) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t)$$

Il valore di questa misura è:

- molto alto quando il termine t compare molte volte in un numero ristretto di documenti, che hanno quindi un alto potere discriminante. Infatti in questo caso entrambi i fattori (tf e idf) sono alti.
- intermedio quando il termine t compare poche volte nel documento d (e quindi tf è basso) oppure quando t compare pressoché in tutti i documenti: significa che t ha un basso potere discriminante e idf sarà basso.
- molto basso quando il termine t compare in praticamente tutti i documenti: idf, che è una misura in scala logaritmica, sarà prossima allo zero.

A questo punto, possiamo vedere un documento d come un vettore $\vec{V}(d)$ in uno spazio k dimensionale, dove k rappresenta il numero di termini distinti presenti nella collezione. Tale vettore ha k componenti: per ciascuna, il peso sarà dato dalla fun-

zione tf-idf. La rappresentazione vettoriale è fondamentale per la classificazione e il raggruppamento (clustering) di documenti.

Definiamo il grado di similarità tra due documenti come il coseno dell'angolo che si forma tra le loro rappresentazioni vettoriali (*cosine similarity*), compensando lo squilibrio dato dalla diversa lunghezza dei vari documenti:

$$\text{Similarity}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} = \vec{v}(d_1) \cdot \vec{v}(d_2)$$

Nella formula, il numeratore è il prodotto scalare tra i due vettori e il denominatore è il prodotto delle loro lunghezze euclidee. L'effetto è sostanzialmente di normalizzazione, riducendo $\vec{V}(d_1)$ e $\vec{V}(d_2)$ ai rispettivi versori $\vec{v}(d_1)$ e $\vec{v}(d_2)$ per evitare di privilegiare i documenti più lunghi.

Applicando questi concetti al nostro caso concreto, consideriamo un documento come l'insieme dei tweet pubblicati da un certo utente. Tramite la classe `CountVectorizer` del framework `sklearn` produciamo una rappresentazione matriciale che indica per ogni documento il numero di occorrenze di ciascun possibile token. A seguire, tramite la classe `TfidfTransformer` trasformiamo il numero grezzo di occorrenze nella misura tf-idf, valutando il peso di ciascun termine in ogni documento. Quest'ultimo insieme di dati costituisce le feature che useremo per la classificazione.

5.2 UN PRIMO TENTATIVO DI CLASSIFICAZIONE

Per identificare automaticamente persone reali a potenziale rischio di binge drinking, è stato effettuato un primo tentativo di classificazione, a seguito del processo descritto nel Capitolo 3 in cui erano stati separati i profili appartenenti a persone reali dai restanti. Lo schema complessivo è illustrato in Figura 4.

I dati costituenti il dataset di training sono stati annotati manualmente da alcuni esperti dell'équipe di psichiatria, come già spiegato nella Sezione 3.3.3. Riassumendo, dagli $N = 494$ utenti iniziali sono stati esclusi i profili non etichettati come appartenenti a persone fisiche, rimanendo con $N' = 180$ profili di utenti fisici etichettati manualmen-

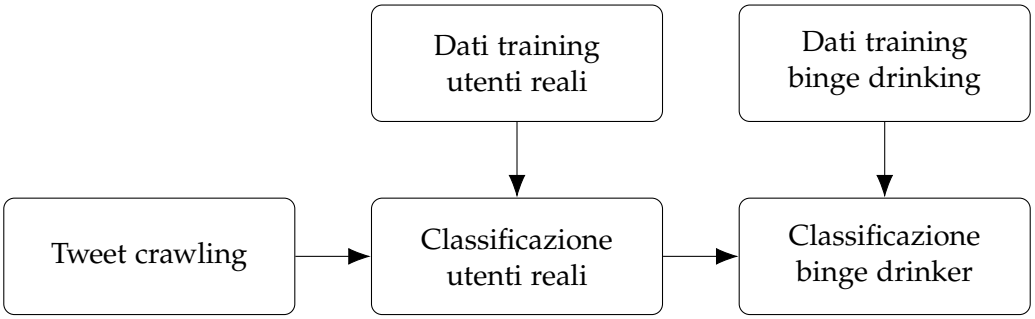

Figura 4: Pipeline per la classificazione degli utenti a rischio binge drinking.

te. Di questi, 45 sono stati ulteriormente identificati come potenzialmente a rischio binge drinking.

Si è quindi proceduto alla valutazione di un classificatore random forest, in analogia a quanto già visto nel terzo capitolo, tramite 5-fold cross validation. Lo squilibrio tra i campioni delle due classi è stato risolto assegnando alle classi peso proporzionale alla loro dimensione.

I risultati sono stati deludenti: il classificatore spesso non prevede alcun utente a rischio binge drinking rispetto a quelli identificati come potenziali binge drinker dagli esperti e la curva ROC mostra che la predizione si avvicina alla totale aleatorietà (cfr. Figura 6).

Una delle possibili ragioni risiede nel fatto che il materiale su cui stiamo costruendo le feature, i tweet correlati all'alcol presenti nel dataset, è davvero esiguo: disponiamo solo di poco più di due tweet per autore (cfr. Sezione 3.1). È impensabile riuscire a determinare una sorta di "lessico comune", utilizzato da utenti potenzialmente a rischio binge drinking, con così poca produzione testuale.

5.3 UN SECONDO TENTATIVO DI CLASSIFICAZIONE

Per migliorare i risultati prodotti dall'approccio descritto nella Sezione 5.2, un ulteriore tentativo deve quindi passare attraverso la raccolta di una quantità maggiore tweet. È stato sviluppato un crawler ad-hoc per scaricare l'intera cronologia di tweet degli uten-

ti classificati manualmente come potenziali binge drinkers nel training set. In totale ne sono stati recuperati 86 204, per una media di oltre 1 900 tweet per utente.

La nuova pipeline di elaborazione su cui si basa questo secondo tentativo di classificazione è illustrata in Figura 5.

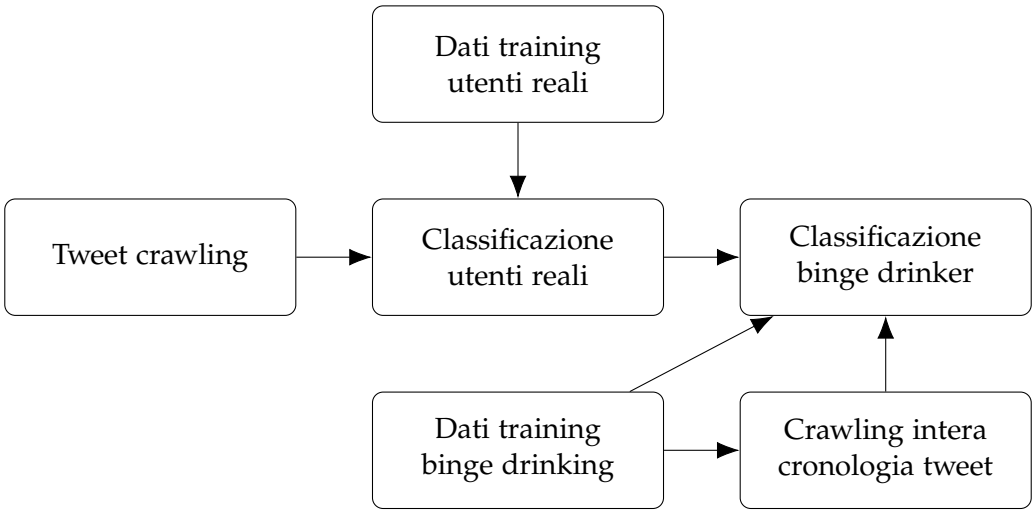

Figura 5: Pipeline migliorata per la classificazione degli utenti a rischio binge drinking.

La curva ROC ottenuta con questo secondo tentativo è illustrata in Figura 7; i valori delle diverse metriche sono riportati invece nella Tabella 8. Il valore di accuratezza raggiunge 0.75 ± 0.06 ma è fortemente influenzato dalla numerosità del campione della classe negativa. I valori di precision, recall e F1 mostrano in modo lampante l'insufficienza di questo approccio per identificare con precisione possibili binge drinker.

	Accuracy	Precision	Recall	F1 score
Random Forest	0.75 ± 0.06	0.4 ± 0.75	0.11 ± 0.2	0.17 ± 0.3

Tabella 8: Misure per la valutazione del classificatore Random Forest applicato all'individuazione di potenziali binge drinker.

Si possono ipotizzare alcune ragioni per cui l'individuazione di simili profili abbia un così basso tasso di successo:

- La scarsità di dati annotati. Essendo questo uno dei primi esperimenti per l'identificazione di utenti a rischio binge drinking su Twitter, non esistono dataset

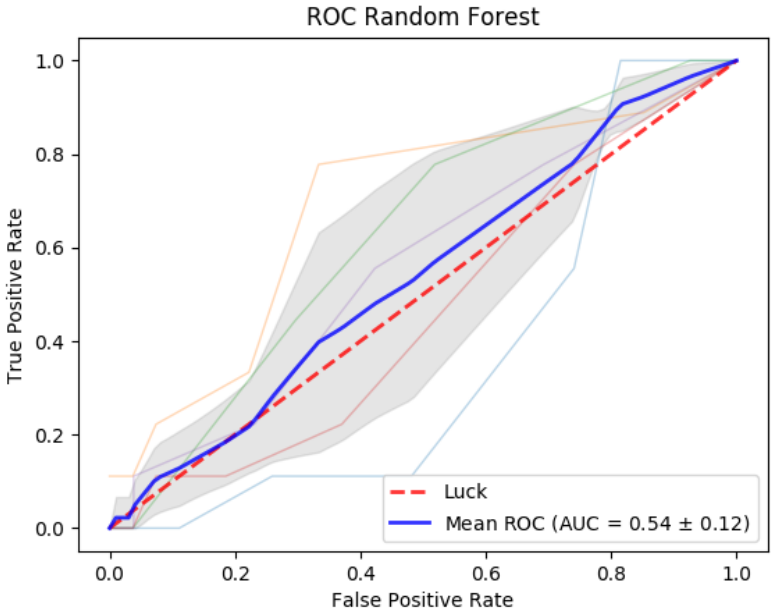

Figura 6: Valore medio AUC e curva ROC del classificatore random forest per l'identificazione di potenziali binge drinker con il primo approccio.

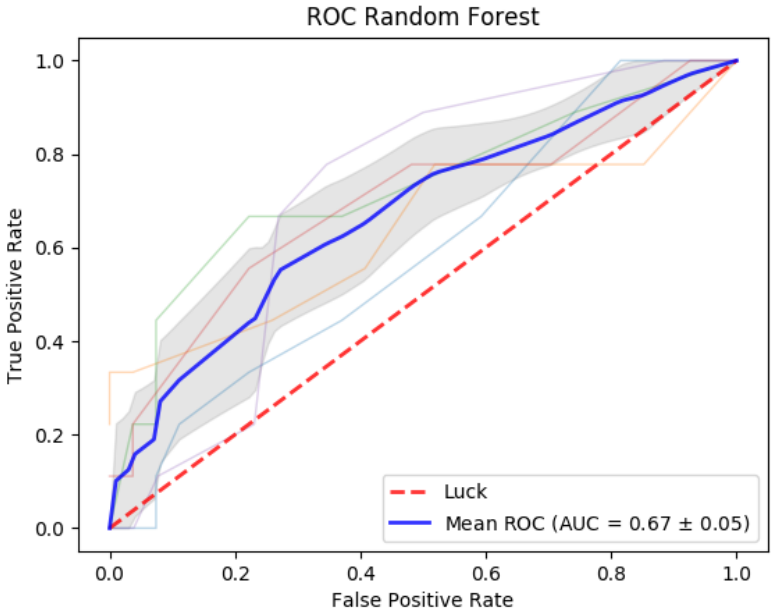

Figura 7: Valore medio AUC e curva ROC del classificatore random forest per l'identificazione di potenziali binge drinker con il secondo approccio.

pubblici già annotati, che possono svolgere la duplice funzione di training set e benchmark rispetto ad altri approcci. Per i fini di questa ricerca sono stati annotati manualmente circa 500 profili, un numero comunque insufficiente per le metodologie di apprendimento automatico.

- Lo scarso utilizzo di Twitter da parte della fascia giovanile. Già in [Duggan and Brenner 2013] si evidenziava come Twitter fosse meno utilizzato di Facebook e Instagram tra gli adolescenti e i giovani. Negli ultimi anni questo trend si è ulteriormente rafforzato e ricerche di questo tipo dovrebbero tenere in considerazione l'analisi di media alternativi, per esempio fotografie prese da Instagram.
- La naturale scarsa propensione a parlare di momenti privati o tra amici, come può esserlo il consumare alcolici, tramite messaggi pubblici.
- La natura stessa dei tweet, che sono una fonte di informazioni spesso fortemente rumorosa e su cui le performance degli approcci tradizionali in ambito NLP sono compromesse in modo importante [Ritter et al. 2011].

Nonostante i problemi evidenziati, questo lavoro costituisce un tentativo iniziale per la classificazione di utenti Twitter a potenziale rischio binge drinking ed è un primo passo su cui basare possibili ulteriori sviluppi in questa direzione.

BIBLIOGRAFIA

Bartoli, F., Carretta, D., Crocamo, C., Schivalocchi, A., Brambilla, G., Clerici, M., and Carrà, G. (2014). Prevalence and correlates of binge drinking among young adults using alcohol: a cross-sectional survey. *BioMed research international*, 2014.

Bian, J., Topaloglu, U., and Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM.

Carrà, G., Crocamo, C., Schivalocchi, A., Bartoli, F., Carretta, D., Brambilla, G., and Clerici, M. (2015). Risk estimation modeling and feasibility testing for a mobile ehealth intervention for binge drinking among young people: The d-arianna (digital-alcohol risk alertness notifying network for adolescents and young adults) project. *Substance abuse*, 36(4):445–452.

Chen, E. E. and Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological methods*, 21(4):458.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824.

Dredze, M. (2012). How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84.

Duggan, M. and Brenner, J. (2013). *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project Washington, DC.

Guo, D. and Chen, C. (2014). Detecting non-personal and spam users on geo-tagged twitter network. *Transactions in GIS*, 18(3):370–384.

Igawa, R. A., Barbon Jr, S., Paulo, K. C. S., Kido, G. S., Guido, R. C., Júnior, M. L. P., and da Silva, I. N. (2016). Account classification in online social networks with lbca and wavelets. *Information Sciences*, 332:72–83.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., and Gonzalez, G. (2014). Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Rushe, D. (2018). Twitter reports profit for second quarter in a row and adds 6m new users. *The Guardian*.

Thomas, K., Grier, C., Song, D., and Paxson, V. (2011). Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM.

Tsukayama, H. (2013). Twitter turns 7: Users send over 400 million tweets per day. *The Washington Post*, 21.

Wechsler, H., Dowdall, G. W., Davenport, A., and Rimm, E. B. (1995). A gender-specific measure of binge drinking among college students. *American journal of public health*, 85(7):982–985.

RINGRAZIAMENTI

Il primo pensiero va ai miei genitori, che mi hanno concesso il privilegio di dedicarmi esclusivamente agli studi. Non è retorica affermare che tutto questo è stato concretamente possibile solo grazie a loro.

Il secondo doveroso ringraziamento va alla Prof.ssa Pasi e al Dott. Viviani, che mi hanno guidato nell'affrontare un argomento in gran parte inesplorato e socialmente rilevante come è quello del binge drinking: cammino bello e difficile al tempo stesso. Come accade spesso, le cose belle sono difficili e le cose difficili sono belle.

Infine, meritano un "grazie" sincero tutti gli amici, stretti o visti solo qualche volta, di lungo corso o conosciuti da poco, che ho incontrato in questi anni. Collaborare, scambiarsi idee e opinioni, arrabbiarsi e gioire, confrontarsi e scontrarsi: è stata una fonte preziosa di arricchimento di cui conserverò a lungo i frutti.