Estrazione pratica di informazioni rilevanti per disastri da Social media

Muhammad Imran *
Università degli Studi di Trento
imran@disi.unitn.it

Shady Elbassuoni
Università americana di Beirut
se58@aub.edu.lb

Carlos Castillo
Qatar Computing
Istituto di ricerca
chato@acm.org

Fernando Diaz
Microsoft Research
fdiaz@microsoft.com

Patrick Meier
Qatar Computing
Istituto di ricerca
pmeier@qf.org.qa

ASTRATTO

Durante i periodi di catastrofi gli utenti online generano una quantità significativa di dati, alcuni dei quali sono estremamente preziosi per gli sforzi di soccorso. In questo articolo, studiamo la natura dei contenuti dei social media generati durante due di ff erenti disastri naturali. Formiamo anche un modello basato su campi casuali condizionali per estrarre informazioni preziose da tali contenuti. Valutiamo le nostre tecniche sui nostri due set di dati attraverso una serie di esperimenti progettati con cura. Testiamo anche i nostri metodi su un set di dati non disastroso per dimostrare che il nostro modello di estrazione è utile per estrarre informazioni dai contenuti generati socialmente in generale.

Categorie e descrittori di soggetto

I.2.7 [**Elaborazione del linguaggio naturale]:** Analisi del testo

Parole chiave

Social media; Filtro delle informazioni; Estrazione delle informazioni

1. INTRODUZIONE

Le piattaforme di microblogging sono diventate un modo importante per condividere informazioni sul Web, specialmente durante eventi critici come i disastri naturali e causati dall'uomo. Negli ultimi anni, Twitter è stato utilizzato per diffondere notizie su vittime e danni, sforzi di donazione e avvisi, comprese informazioni multimediali come video e foto [1, 3]. Data l'importanza dei tweet sull'argomento per la consapevolezza della situazione critica in termini di tempo, le comunità colpite da catastrofi e i soccorritori professionali possono trarre vantaggio dall'utilizzo di un sistema automatico per estrarre le informazioni rilevanti da Twitter.

Proponiamo un metodo in due fasi per l'estrazione delle informazioni relative al disastro: (i) classificazione dei tweet e (ii) estrazione dai tweet. La fase di classificazione si basa sul nostro lavoro precedente [8]; la fase di estrazione è il fulcro di questo documento. Entrambi i passaggi vengono eseguiti utilizzando software libero o-the-shelf [6, 7], ottenendo un sistema facile da implementare e che secondo i nostri esperimenti ha buone prestazioni.

* Lavoro svolto mentre l'autore era al QCRI.

WWW 2013 Companion, 13-17 maggio 2013, Rio de Janeiro, Brasile. ACM 978-1-4503-2038-2 / 13/05.

Il resto del lavoro è organizzato come segue. La Sezione 2 descrive il nostro metodo di estrazione delle informazioni, che viene valutato nella Sezione 3. La Sezione 4 mostra che il nostro metodo può essere applicato anche in contesti non disastrosi. Nella sezione 5, delineiamo brevemente i lavori correlati e concludiamo nella sezione 6.

2. DESCRIZIONE DEL NOSTRO APPROCCIO

Questa sezione descrive le fasi di classificazione ed estrazione del nostro metodo. Per chiarezza di esposizione e concretezza, iniziamo descrivendo i dataset che utilizziamo.

2.1 Set di dati

Utilizziamo due set di dati relativi alle recenti emergenze:

Joplin 2011: 206.764 tweet raccolti durante il tornado che ha colpito Joplin, Missouri (USA) il 22 maggio 2011. I ricercatori dell'Università del Colorado a Boulder hanno raccolto il set di dati tramite l'API di Twitter utilizzando l'hashtag [1] # joplin.

Sandy 2012: 140.000 tweet raccolti durante l'uragano Sandy, che ha colpito il nord-est degli Stati Uniti il 29 ottobre 2012. Il set di dati è stato raccolto utilizzando gli hashtag # sandy, #nyc.

2.2 Classificazione

Poiché i messaggi generati durante un disastro sono estremamente vari, un sistema automatico deve iniziare filtrando i messaggi che non contribuiscono a informazioni preziose. Questi includono quelli che sono interamente di natura personale e quelli non rilevanti per la crisi in corso. In particolare, iniziamo

di s • dividendo i messaggi in due classi principali:

- **Personale:** se un messaggio interessa solo il suo autore e la sua cerchia immediata di familiari / amici e non trasmette alcuna informazione utile alle persone che lo fanno non conosco il suo autore.
- **Informativo:** se il messaggio è *Informativo (* di interesse ad altre persone al di fuori della cerchia immediata dell'autore).
- **Altro:** se il messaggio non è correlato al disastro.

Inoltre, si distinguono due tipi di messaggi informativi: diretti, cioè scritti da una persona che è un testimone oculare diretto di ciò che sta accadendo o indiretti, quando il messaggio ripete informazioni riportate da altre fonti.

Una volta rilevati i tweet informativi, li classifichiamo nelle seguenti classi (i dettagli sulla scelta di questa ontologia possono essere trovati in [8]):

[1] Questi hashtag sono per lo più annunciati dalle autorità di gestione delle crisi al momento di un incidente.

Tabella 1: Istruzioni dipendenti dal tipo fornite ai valutatori per la fase di estrazione ed esempio (in grassetto) della parte estratta.

genere	**Istruzioni:** Copia e incolla la parola / frase che ... **Esempio**	
Attenzione o consiglio: tutti	. . . avverte di un potenziale pericolo o consiglia cosa fare	. @NYGovCuomo ordina **chiusura dei ponti di New York.** Solo I ponti di Staten Island non sono stati toccati in questo momento. I ponti devono chiudere entro le 19:00. #Sandy #NYC.
Fonte di informazione: Photos / video	. . . indica di cosa tratta il contenuto di una foto / video	RT @NBCNewsPictures: **Foto dell'incredibile scene abili lasciate sulla scia di #Hurricane # Sandy** http://t.co/09U9L5rW #NYC #NJ
Persone: ritrovate persone scomparse o perse	. . . indica chi manca o è stato trovato	rt @ 911bu ff: aiuto pubblico necessario: **2 ragazzi 2 e 4 scomparsi** quasi 24 ore dopo che si erano separati dalla madre quando l'auto è stata sommersa nel si. #sandy # 911bu ff
Vittime e danni: Infrastruttura	. . . indica una struttura, una strada, un servizio, una linea, ecc. che non funziona o è stata danneggiata	RT @TIME: l'edificio di New York aveva già ricevuto numerosi reclami di costruzione **crollo della gru** http://t.co/7EDmKOp3 # Sabbioso
Vittime e danni: Ferito o morto	. . . indica chi (o quante persone) è stato ferito o morto	Almeno **39** milioni di morti senza corrente all'indomani di Sandy. http://t.co/Wdvz8KK8
Donazioni: denaro / beni / servizi Richieste	. . . indica cosa (denaro, beni, lavoro, servizi gratuiti, ecc.) viene richiesto come donazione	**Occorrono 400 volontari** per le aree che #Sandy ha distrutto.
Donazioni: denaro / beni / servizi O ff ers	. . . indica cosa (denaro, beni, lavoro, servizi gratuiti, ecc.) viene offerto in donazione	Voglio offrirmi volontario **aiutare le vittime dell'uragano Sandy.** Se qualcuno sa come posso essere coinvolto per favore fatemelo sapere!
Persone: Celebrilegami / autorità	. . . nomina una celebrità o un'autorità che reagisce all'evento o visita l'area	**Candidato VP Ryan** partecipa a una raccolta di cibo in Wisconsin per le vittime dell'uragano Sandy. PO-35WE su BitCentral.

- **Attenzione e consiglio:** se un messaggio trasmette / riporta informazioni su un avvertimento o un consiglio su un possibile rischio di incidente.
- **Vittime e danni:** se un messaggio riporta il file informazioni su vittime o danni alle infrastrutture fatto da un incidente.
- **Donazioni** di denaro, beni o servizi: se un messaggio parla di beni o servizi offerti o di cui necessitano le vittime di un incidente.
- **Persone** mancante, trovato o visualizzato: se viene segnalato un messaggio su una persona scomparsa o trovata colpita da un incidente ammaccatura o segnala la reazione o la visita di una celebrità.
- **Fonti di informazione:** se un messaggio punta a informazioni fonti di mazione, foto, video; o menziona un sito web, Stazione televisiva o radiofonica che fornisce un'ampia copertura.
- **Altro:** altri tipi di messaggi informativi.

Come descritto nel nostro lavoro precedente [8], una serie di classificatori multi-etichetta è stata addestrata a classificare automaticamente un tweet in una o più delle classi precedenti. I classificatori bayesiani na¨ive sono usati come implementati in Weka [7]. I nostri classi fi catori utilizzano un ricco set di funzionalità tra cui unigrammi di parole, bigrammi, tag Partof-Speech (POS) e altri. Il nostro set di funzionalità contiene oltre a un set di funzionalità binarie (ad esempio, se un tweet contiene un URL, un'emoticon, un hashtag, ecc.) E funzionalità scalari (come la lunghezza del tweet). I dati di formazione per i nostri classificatori sono stati ottenuti classificando manualmente una serie di tweet utilizzando il crowdsourcing tramite il provider Crowd fl ower [2].

Abbiamo ottenuto circa 2.000 etichette per il set di dati Sandy e circa 4.400 per il set di dati Joplin.

2.3 Estrazione

Una volta che un tweet è stato classi fi cato in una delle classi di cui sopra, le informazioni rilevanti per la classe possono essere estratte per ulteriori analisi. Ad esempio, per un file *vittime e danni* tweet, è possibile identificare il numero di vittime o il nome dell'infrastruttura che è stata danneggiata.

[2] http: //www.crowd fl ower.com

Abbiamo trattato il compito di rilevare le informazioni rilevanti per la classe come un'attività di etichettatura di sequenza. Un tweet è considerato una sequenza di simboli di parole. In un'attività di etichettatura in sequenza, ogni token viene etichettato algoritmicamente come parte di una sottosequenza di informazioni di destinazione o come non correlato a tali informazioni. Nell'esempio del primo tweet nella Tabella 1, i token "chiusura", "of", "NYC" e "bridge" sono etichettati come positivi (parte delle informazioni sul target), mentre il resto dei token è etichettato come negativo. Di seguito è riportato un esempio: si noti che anche il punto (".") È un segno:

. . . ordina la chiusura dei ponti di New York. Solo Staten ...
- + + + + - - -

Usiamo campi casuali condizionali, un algoritmo di etichettatura di sequenze apprese dalla macchina, per il nostro compito [9]. Un campo casuale condizionale (CRF) è un modello probabilistico che, nel nostro compito, prevede l'etichetta di ogni token ("+" o "-") date entrambe le informazioni endogene al token (es. "Token is a number", "token è la parola *ponti* ') così come le informazioni esogene al token (es.' token è preceduto dalla parola

chiusura '). I CRF sono stati applicati con successo in passato ad altre attività di estrazione di informazioni [10].

Usiamo ArkNLP, un'implementazione di CRF e una serie di funzionalità note per essere efficaci per le attività di PNL sui dati di Twitter [6]. In pratica, modifichiamo semplicemente i dati di addestramento di ArkNLP per conformarli a quanto descritto sopra, e li eseguiamo senza ulteriori modi fi che.

Attività di crowdsourcing. Durante l'attività di crowdsourcing per l'estrazione, mostriamo ai valutatori ogni tweet e il tipo (e il sottotipo, se disponibile) determinato durante la fase di classificazione. Usiamo un'istruzione specifica per ogni sottotipo, come elencato nella colonna "istruzione" della Tabella 1.

Agli operatori è stato mostrato un tweet, questa istruzione e un campo di immissione di testo vuoto e gli è stato chiesto di copiare e incollare una parola o una breve frase dal tweet che trasmette le informazioni specificate. Non abbiamo accettato alcun esempio di formazione in cui il segmento estratto dal lavoratore in crowdsourcing non fosse contenuto nel tweet originale.

3. RISULTATI SPERIMENTALI

Metrica. Valutiamo il nostro sistema confrontando il suo output con le risposte fornite dagli esseri umani. Noi **treno** il nostro sistema su una parte delle etichette fornite dall'uomo e **test**
il sistema sulla parte restante. Ci sono due aspetti che misuriamo che sono legati alla sensibilità e alle speci fi che del nostro sistema.

Tasso di rilevamento (analogo alla sensibilità statistica, o richiamo) misura la frazione di esempi in cui gli esseri umani hanno trovato un'informazione rilevante e anche il nostro sistema ha trovato qualcosa, anche se quel qualcosa non è corretto.

Percentuale di successi (analogo a uno meno la specificazione, o precisione) misura la frazione di esempi per i quali il nostro sistema ha trovato qualcosa e quel qualcosa potrebbe essere considerato corretto dagli esseri umani. Consideriamo corretto l'output se si sovrappone in almeno una parola con l'etichetta umana data.

Esempio di metriche. Un esempio può illustrare queste metriche. Supponiamo che l'input e l'output siano i seguenti:

	Ingresso	Produzione
un	C'erano **12** ferito	**<** vuoto**>**
b	UN **il ponte è crollato** ponte	
c	**10 volontari** necessario	necessario

In questo caso, il tasso di rilevamento è del 66%, dato che in due ({ *avanti Cristo}*) dei 3 esempi il nostro sistema ha rilevato qualcosa. L'hit ratio è del 50% dato che solo in uno dei due (*b)*
l'output si sovrappone all'estrazione di destinazione nell'input.

Risultati generali. La tabella 2 mostra i risultati dei nostri vari esperimenti, in cui abbiamo selezionato le classi più grandi che avevamo a disposizione: cautela e consigli, vittime e danni: infrastrutture e donazioni. In generale, e in modo simile agli scambi di precisione-richiamo osservati nei sistemi di recupero delle informazioni, spesso un tasso di rilevamento più elevato è associato a un rapporto di successo inferiore e viceversa.

Ci sono quattro blocchi che studiano diversi scenari. Concentriamoci per ora sulla prima riga di ogni blocco, dove *Treno*
è "Tutti" e *Test* è tutto".

I primi due blocchi misurano le prestazioni del nostro sistema Joplin e sabbioso dati. Il tasso di rilevamento è più alto per Joplin (78%) rispetto a Sabbioso (41%). Il rapporto di successo è anche più alto per Joplin (90%) rispetto a Sabbioso (78%). Ciò indica che il secondo set di dati è più impegnativo per il nostro sistema rispetto al primo. Tuttavia, in entrambi i casi il rapporto di successo è piuttosto alto, indicando che quando il nostro sistema estrae una parte del tweet, spesso è la parte corretta.

Il terzo blocco misura le prestazioni di un ipotetico sistema addestrato sui dati da Joplin, e quindi testato sui dati di Sabbioso. Questo di solito è indicato come un file *adattamento* o *trasferimento* scenario. Possiamo osservarlo rispetto a uno scenario in cui ci alleneremmo sui dati Sabbioso, il tasso di rilevamento scende drasticamente (11% vs 41%), mentre il tasso di successo non è influenzato in modo significativo (78% vs 79%).

Le classi di tweet più colpite sono quelle che forniscono cautela e consigli, che sembrano essere abbastanza specifici per l'evento. D'altra parte, la performance per i tweet relativi alla donazione è quella meno influenzata tra le tre classi, indicando che le parole e le frasi usate per descrivere non variano tanto quanto per le altre classi da un evento all'altro.

Nel quarto blocco, consideriamo uno scenario di adattamento in cui una quantità limitata di nuovi dati (da Sabbioso) è incorporato nella formazione. Questo simula un caso in cui

Tabella 2: Prestazioni della fase di estrazione delle informazioni per diverse con fi gurazioni di training e test set. "Tutto" significa nessuna distinzione tra categorie. La seconda e la quarta colonna mostrano il numero di tweet rispettivamente nei dati di addestramento e di test.

Allenati sul 66% di Joplin, prova sul 33% di Joplin						
Treno		Test		Rilevato	Det. Vota	Percentuale di successi
Tutti	338	Tutti	169	131	78%	90%
Tutti	338	CIRCA	130	109	84%	93%
Tutti	338	Infra.	4	3	75%	33%
Tutti	338	Dona.	34	25	74%	92%
CIRCA	260	CIRCA	130	118	91%	95%
Infra.	10	Infra.	4	1	25%	0%
Dona.	69	Dona.	34	16	47%	81%
Allenati sul 66% di Sandy, prova sul 33% di Sandy						
Treno		Test		Rilevato	Det. Vota	Percentuale di successi
Tutti	397	Tutti	198	82	41%	79%
Tutti	397	CIRCA	69	27	39%	74%
Tutti	397	Infra.	93	71	76%	83%
Tutti	397	Dona.	35	23	66%	83%
CIRCA	139	CIRCA	69	26	38%	85%
Infra.	187	Infra.	93	50	54%	80%
Dona.	72	Dona.	35	12	34%	83%
Allenati sul 100% di Joplin, prova sul 100% di Sandy						
Treno (Joplin)		Test (Sandy)		Rilevato	Det. Vota	Percentuale di successi
Tutti	507	Tutti	595	66	11%	78%
Tutti	507	CIRCA	208	4	2%	100%
Tutti	507	Infra.	280	24	9%	71%
Tutti	507	Dona.	107	38	36%	82%
CIRCA	390	CIRCA	208	2	1%	100%
Infra.	14	Infra.	280	44	16%	73%
Dona.	103	Dona.	107	52	49%	90%
Allenati su 100% Joplin + 10% di Sandy, Prova il 90% di Sandy						
Treno (Joplin +)		Test (Sandy-)		Rilevato	Det. Vota	Percentuale di successi
Tutti	568	Tutti	534	112	21%	81%
Tutti	568	CIRCA	187	9	5%	100%
Tutti	568	Infra.	251	64	25%	80%
Tutti	568	Dona.	96	39	41%	79%
CIRCA	411	CIRCA	187	18	10%	71%
Infra.	43	Infra.	251	106	42%	83%
Dona.	114	Dona.	96	46	48%	89%

aspettiamo alcune ore prima di generare un output, in modo da ottenere alcuni esempi etichettati sul nuovo evento. Le prestazioni sono superiori rispetto al caso precedente, con un tasso di rilevamento del 21% e un tasso di successo dell'81%.

Quest'ultimo risultato mostra che possiamo migliorare in modo incrementale il nostro modello per funzionare meglio ogni volta che abbiamo bisogno di usarlo in un nuovo disastro.

Risultati dettagliati. In ogni blocco, la prima riga riporta il tasso di rilevamento e il rapporto di successo durante l'allenamento *un unico modello*
su tutti i tweet nel nostro set di allenamento e lo testiamo su tutti i tweet nel nostro set di test indipendentemente dalle rispettive classi dei tweet. Nelle tre righe successive disaggreghiamo questa impostazione per ogni classe nella parte di test. Infine, nelle ultime tre righe mostriamo le prestazioni quando ci alleniamo *tre di ff erenti modelli,* uno per ogni classe e provalo solo sui tweet della stessa classe.

I risultati indicano che i modelli specifici per classe possono portare a miglioramenti nelle prestazioni per alcune classi ma non per altre. I modelli specifici per classe sono particolarmente utili per la classe di attenzione e consiglio dei tweet e producono miglioramenti nel tasso di rilevamento per sabbioso set di dati nel caso di tweet relativi a donazioni. Non ci sono guadagni consistenti per i tweet relativi ai danni alle infrastrutture, tranne durante la formazione Joplin e testare Sabbioso.

4. GENERALIZZAZIONE DEGLI ALTRI EVENTI

Un approccio robusto dovrebbe generalizzarsi a una varietà di scenari, inclusi eventi non correlati a disastri. In questa sezione discutiamo brevemente una serie di esperimenti su un set di dati non disastroso corrispondente a una partita sportiva. Il set di dati, che consiste di 72.000 tweet, è stato raccolto utilizzando *Twitter Streaming API* utilizzando # cricket, #indvspak, #indvpk hashtag durante una partita di cricket tra Pakistan e India il 6 gennaio,

2013.

Attività di crowdsourcing. Etichettiamo i dati utilizzando la stessa procedura degli altri nostri set di dati. Nella prima attività, che comprendeva 2.000 tweet unici, abbiamo chiesto ai lavoratori di etichettare un tweet individuale in (i) separare i tweet informativi da quelli personali e (ii) per un tweet informativo specificare quali informazioni trasmette.

Abbiamo utilizzato sei classi che dipendono dal dominio e corrispondono a eventi durante una partita di cricket: limite, punteggio, over, espulsione, palla e altro [3]. Nella seconda attività, che comprendeva 631 tweet informativi, ai lavoratori sono stati presentati il tipo e il sottotipo di un tweet e gli è stato chiesto di copiare e incollare una parola o una breve frase utilizzando un'istruzione dipendente dal tipo.

Risultati sperimentali. La tabella 3 mostra i risultati di vari esperimenti su questo set di dati. Le prime due righe sono scenari in cui viene creato un singolo modello e le restanti corrispondono a più modelli specifici della classe. Quando ci si allena sull'intero set di allenamento e si esegue il test sull'intero set di test, si osserva un tasso di rilevamento relativamente basso. Questo può essere migliorato se incorporiamo esempi in cui più di un tipo di informazione è presente in un dato tweet, come mostrato nella seconda riga. Possiamo anche vedere miglioramenti significativi nel rapporto di successo per tutti i modelli specifici della classe.

Tabella 3: Risultati con dati sul cricket.

	Formazione casi	Test casi	Rilevamento Vota	Colpire rapporto
Tutti	321	161	43%	95%
Tutto (più etichette)	321	161	51%	95%
Punto	129	66	65%	98%
Altro	100	51	76%	92%
Licenziamento	63	31	81%	88%
Confine	18	8	88%	100%
Palla	6	3	100%	100%
Al di sopra di	5	2	50%	100%

5. LAVORI CORRELATI

Durante le emergenze le piattaforme di social media come Facebook e Twitter distribuiscono informazioni aggiornate sulla consapevolezza situazionale (ad es. Danni, causalità, ecc.) In tutte le forme (ad es. Foto, video, ecc.) [2, 3]. Cameron et al. [4] descrivono una piattaforma per la consapevolezza delle situazioni di emergenza, che rileva gli incidenti utilizzando il rilevamento di parole chiave burst e classifica i tweet interessanti utilizzando un classificatore SVM. Tuttavia, l'identificazione di messaggi informativi sull'argomento e l'estrazione di informazioni utilizzabili pone serie sfide a causa della natura rumorosa e non strutturata dei dati di Twitter. La maggior parte dei lavori precedenti erano basati su metodi di apprendimento automatico standard che in genere si formavano su testi di notizie formali e funzionavano male per una fonte estremamente informale come Twitter [5].

In questo articolo abbiamo utilizzato l'approccio classi fi cazione-estrazione presentato nel nostro lavoro precedente [8], adattandoci in modo semplice e in modo semplice la parte del discorso specifica di Twitter associa ArkNLP al nostro compito [6].

[3] http://en.wikipedia.org/wiki/Cricket

6. CONCLUSIONI E LAVORI FUTURI

Abbiamo presentato un sistema pratico in grado di estrarre dai tweet informazioni rilevanti per il disastro. Secondo numerosi esperimenti su due diversi set di dati, il nostro approccio è in grado di rilevare dal 40% all'80% dei tweet contenenti questo tipo di informazioni e generare un output corretto dall'80% al 90% delle volte.

Questa estrazione a livello di tweet è secondo noi la chiave per poter estrarre informazioni affidabili di alto livello. Osservare, ad esempio, che un gran numero di tweet in luoghi simili segnalano che la stessa infrastruttura è stata danneggiata, può essere un forte indicatore del fatto che è davvero così.

Si prega di contattare gli autori per domande sulla disponibilità dei dati.

Ringraziamenti. Un sincero ringraziamento a Kate Starbird e Project EPIC presso l'Università di Boulder, Colorado, per aver condiviso i tweet-id del Joplin set di dati.

7. RIFERIMENTI

[1] Cynthia D. Balana. Social media: strumento principale in risposta ai disastri, 2012.

[2] Fredrik Bergstrand e Jonas Landgren. Informazione condivisione utilizzando video in diretta nel lavoro di risposta alle emergenze. In *Proc. di ISCRAM.* Citeseer, 2009.

[3] Heather Blanchard, Andy Carvin, Melissa Elliott Whitaker e Merni Fitzgerald. Il motivo per integrare la risposta alle crisi con i social media. Libro bianco, Croce Rossa americana, 2012.

[4] Mark A Cameron, Robert Power, Bella Robinson e Jie Yin. Consapevolezza delle situazioni di emergenza da Twitter per la gestione delle crisi. In *Proc. di WWW,* pagine 695–698. ACM, 2012.

[5] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau e Mark Dredze. Annotazione di entità con nome nei dati di Twitter con crowdsourcing. In *Proc. di HLT,* pagine 80–88, 2010.

[6] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Je ff rey Flanigan e Noah A. Smith. Tagging delle parti del discorso per Twitter: annotazioni, funzionalità ed esperimenti. In *Proc. di HLT,* pagine 42–47, Stroudsburg, PA, USA, 2011.

[7] Mark Hall, Eibe Frank, Geo ff rey Holmes, Bernhard Pfahringer, Peter Reutemann e Ian H. Witten. Il software di data mining weka: un aggiornamento. *SIGKDD Explor. Newsl.,* 11 (1): 10-18, novembre 2009.

[8] Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz e Patrick Meier. Estrazione di pepite di informazioni da messaggi relativi a disastri nei social media. In *ISCRAM,* Baden-Baden, Germania, 2013.

[9] John La ff erty, Andrew McCallum e Fernando Pereira. Campi casuali condizionali: modelli probabilistici per la segmentazione e l'etichettatura dei dati di sequenza. In *Proc. di ICML,* pagine 282–289, 2001.

[10] Fuchun Peng e Andrew McCallum. Informazione estrazione da documenti di ricerca utilizzando campi casuali condizionali. *IP&M,* 42 (4): 963 - 979, 2006.