Sii-Mobility

Supporto di Interoperabilità Integrato per i Servizi al Cittadino e alla Pubblica Amministrazione

Trasporti e Mobilità Terrestre, SCN_00112

Deliverable ID: DE4.7a
Titolo: Servizio di monitoraggio Social Media, e di valutazione del Crowdsourcing

Data corrente	12-01-2017
Versione (solo il responsabile puo' cambiare versione	0.3
Stato (draft, final)	Final
Livello di accesso (solo consorzio, pubblico)	Public
WP	WP4
Natura (report, report e software, report e HW..)	Report e software
Data di consegna attesa	M12, Dicembre 2016
Data di consegna effettiva	12-01-2017
Referente primario, coordinatore del documento	UNIFI: DISIT
Contributor	Gianni Pantaleo, gianni.pantaleo@unifi.it Imad Zaza, imad.zaza@unifi.it
Coordinatore responsabile del progetto	Paolo Nesi, UNIFI, paolo.nesi@unifi.it

Sommario

1 Introduzione ed obiettivi

1.1 Acronimi

TV	Twitter Vigilance
OSN	Open Social Network

1.2 Contesto

Il progetto di Twitter Vigilance istituito al Disit presso l'Università di Firenze si propone proprio di esaminare i messaggi del suddetto social network per fornire analisi in "tempo reale" di tutti quei fenomeni che hanno bisogno di un monitoraggio costante e di conclusioni celeri ed efficaci: previsioni di eventi tramite lo studio delle ricorrenze e non solo, descrizione di eventi catastrofici, avvisare il prima possibile di eventi metereologici improvvisi e violenti, monitorare le smart cities etc…

Ma cosa è nel dettaglio questo progetto e come si possono effettuare tutte queste estrapolazioni di informazioni dai dati?

Per le suddette caratteristiche di Twitter, esso può essere considerato un microblog con aspetti sociali e il suo "prodotto" giornaliero, vale a dire tutti i messaggi scritti nell'arco di 24 ore, sono a tutti gli effetti dei Big Data, con tutte le conseguenze che la loro acquisizione, gestione e utilizzo per analisi (in tempo reale) implica. L'obiettivo finale è quello di usare i dati per creare un modello matematico per previsione, diagnosi preventiva/precoce: è proprio la grande quantità di dati che dà valore a queste analisi, perché un esiguo numero di utilizzatori e/o un piccolo numero di tweets rende l'analisi troppo "di nicchia" per poter assumere un valore ad ampio spettro.

I principali elementi considerati e conteggiati sono il numero di tweet, retweet, follower, commenti e tutti quei parametri che possono essere indicizzati: sono proprio questi i dati che la suddetta Twitter Vigilance presente al Disit raccoglie e predispone per l'utilizzo da parte degli utenti della piattaforma.

Il processo di studio e estrapolazione di informazioni inizia con la creazione di uno o più canali di ricerca e monitoraggio con parole chiave scelte ad hoc, per proseguire con l'analisi e la visualizzazione di questi dati (volume, attività, parole usate, sentiment analysis etc) e infine il modello per la creazione delle previsioni, avvisi e altro.

Gli strumenti a disposizione all'interno del portale sono tre: un tool principale, un tool per la ricerca in tempo reale e un tool per la ricerca avanzata, a cui si aggiunge la capacità di creare vere e proprie campagne di monitoraggio e analisi a lungo raggio su eventi di primaria importanza e "portata" che rientrano direttamente nel mondo dei Big Data Analytics.

Il primo tool è comprensivo delle funzioni sovra esposte, mentre il secondo, come suggerisce il nome, si usa per la diagnosi di situazioni impellenti e precoci, come eventi atmosferici o spettacoli televisivi: il terzo tool, invece, si utilizza quando la ricerca deve essere affinata con l'utilizzo di connettori, incrociando più parole chiave o ricerche.

L'utilizzo e le potenzialità di questo metodo di ricerca sembrano importanti e notevoli, come dimostrano l'interesse e lo sviluppo continuo. Per i nostri scopi, in questo lavoro, abbiamo analizzato una grande mole di dati cercando, fra le varie cose, di estrapolare attraverso la sentiment analysis le opinioni riguardo eventi collegati con il mondo Uber ma anche di prevedere affluenze e interessi futuri nei confronti dell'app californiana, sia in termini di utilizzatori totali che di frequenza (picchi, ricorsività settimanale/mensile etc).

1.3 Obiettivi

L'obiettivo di questo documento è dare evidenza che il sistema di acquisizione dati Social media per il monitoraggio del sistema di trasporti è stato attivato e come. Il sistema una volta attivato ha necessità di collezionare dati per alcuni mesi, per arrivare a poter produrre valutazioni statistiche significative. Su tali dati e pertanto nelle prossime versioni di questo stesso documento saranno presentate anche le valutazioni statistiche.

Il documento è stato redatto in lingua inglese per dare maggiore visibilità internazionale al lavoro svolto e poter utilizzare come reference.

2 User Manual

When you enter the front end by unauthenticated or authenticated user the first page that appears is the page "Channel Statistics": This page shows the list of "public" channels both in the form of a graph and in table form (Figure 12). In the column Detail there are two buttons: the first on the left makes access to the detail page of the selected channel, the second is disabled and in the future will show the NLP analysis of tweet of the channel.

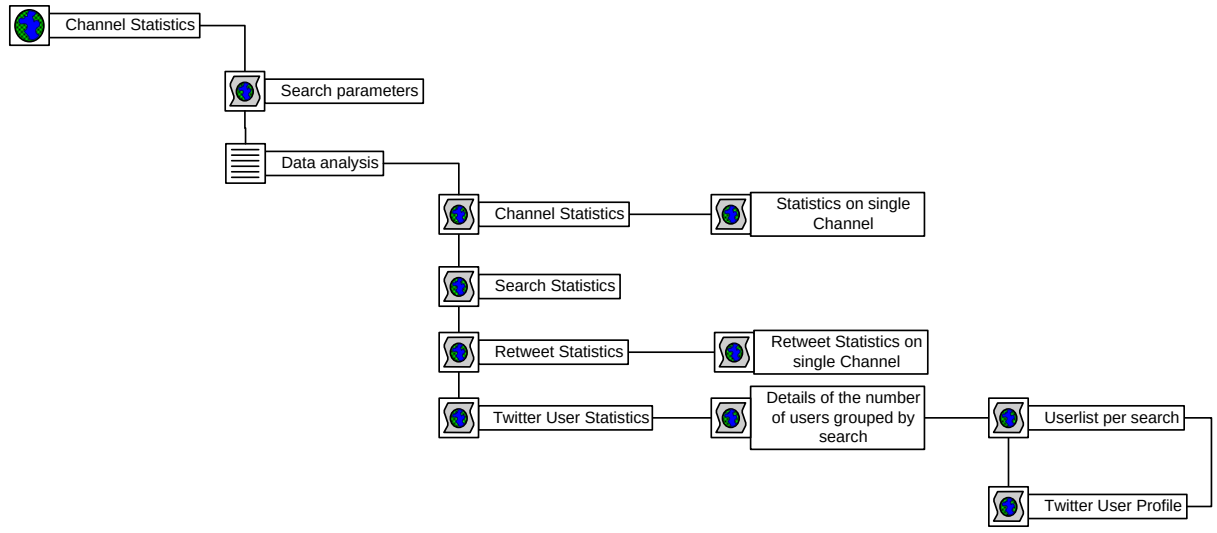
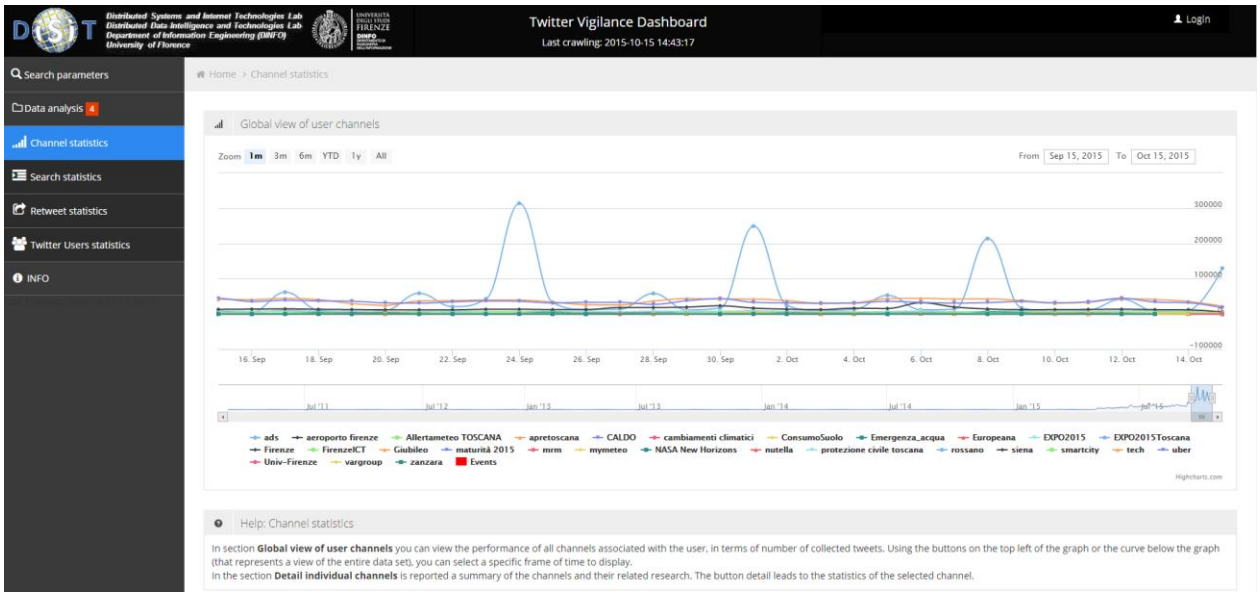


Figure 1: Structure of User Frontend



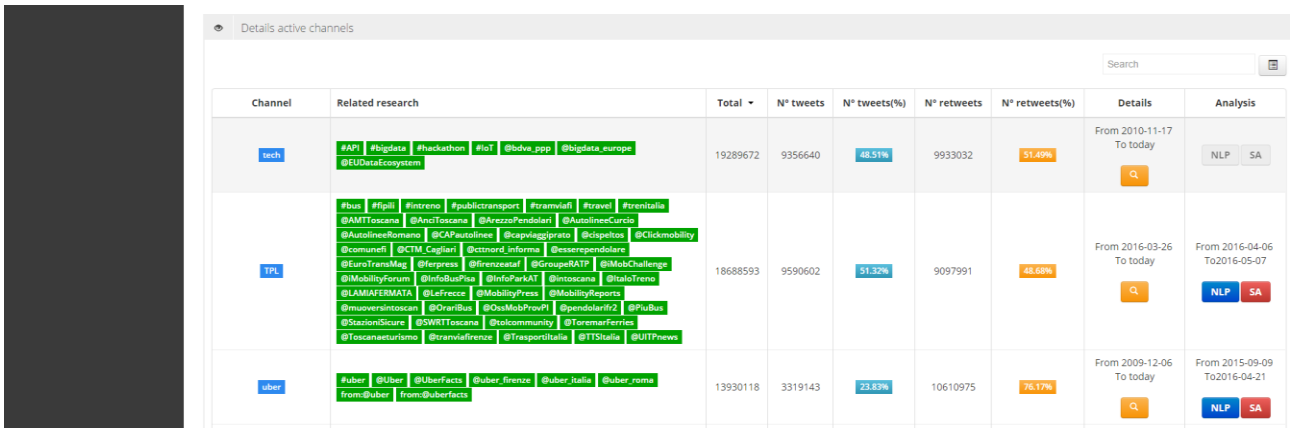


Figure 1: Channel statistics page

The graph shown in Figure 2 shows the number of tweets per day for each channel associated with the user.

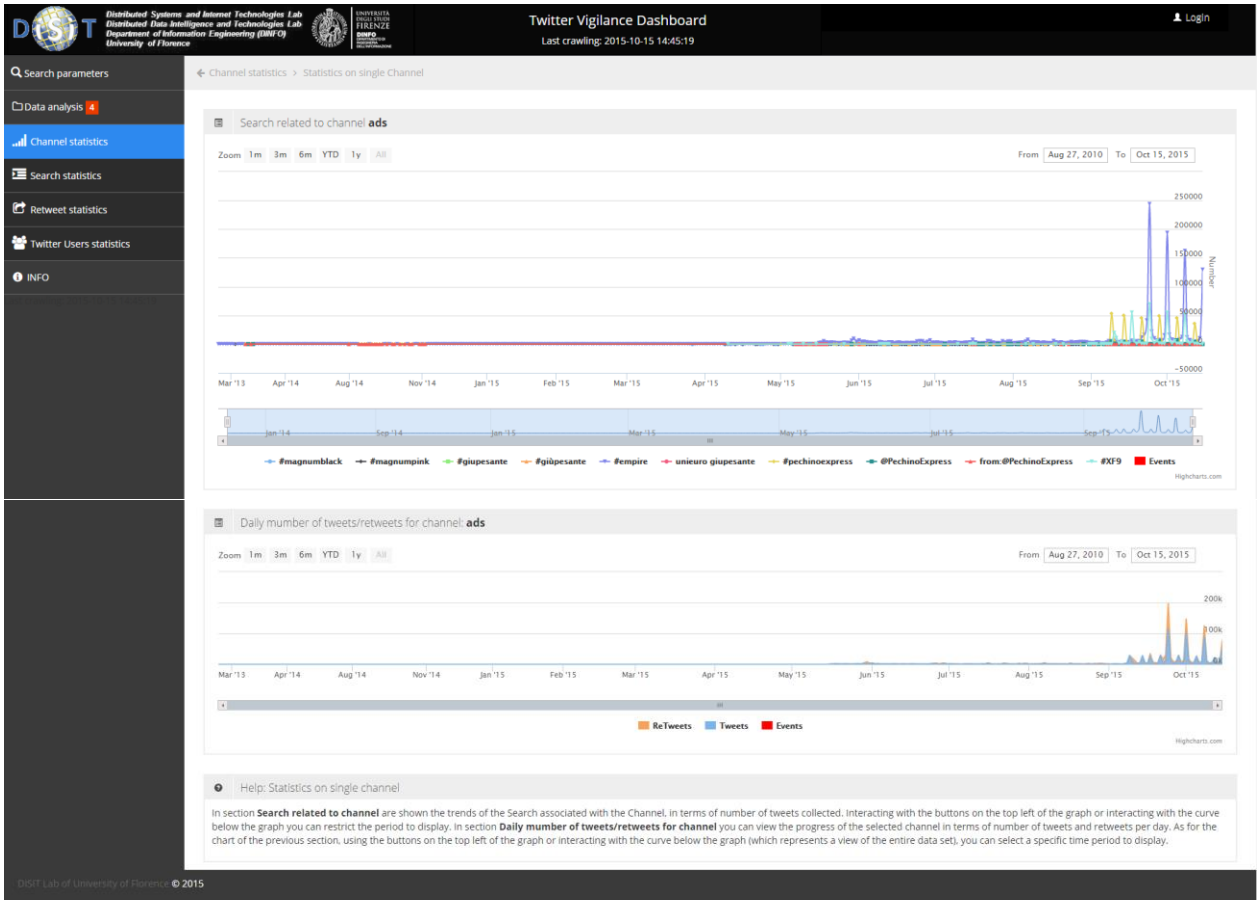


Figure 3: Statistics on single Channel page

Details page of the channel, Figure, shows two graphs: the top one shows the number of tweets per day for each search associated with the channel, the second shows the number of tweets and retweets per day for the entire channel.
For the display of this graph was prepared the table "chart_twitter" in which are stored the necessary data by a separate process which will be described later. This table has updated at regular intervals by adding only the values for the downloaded messages since the last update.

2.1 Search Statistics

On this page (Figure) you can see a histogram that shows the total number of messages for each channel.

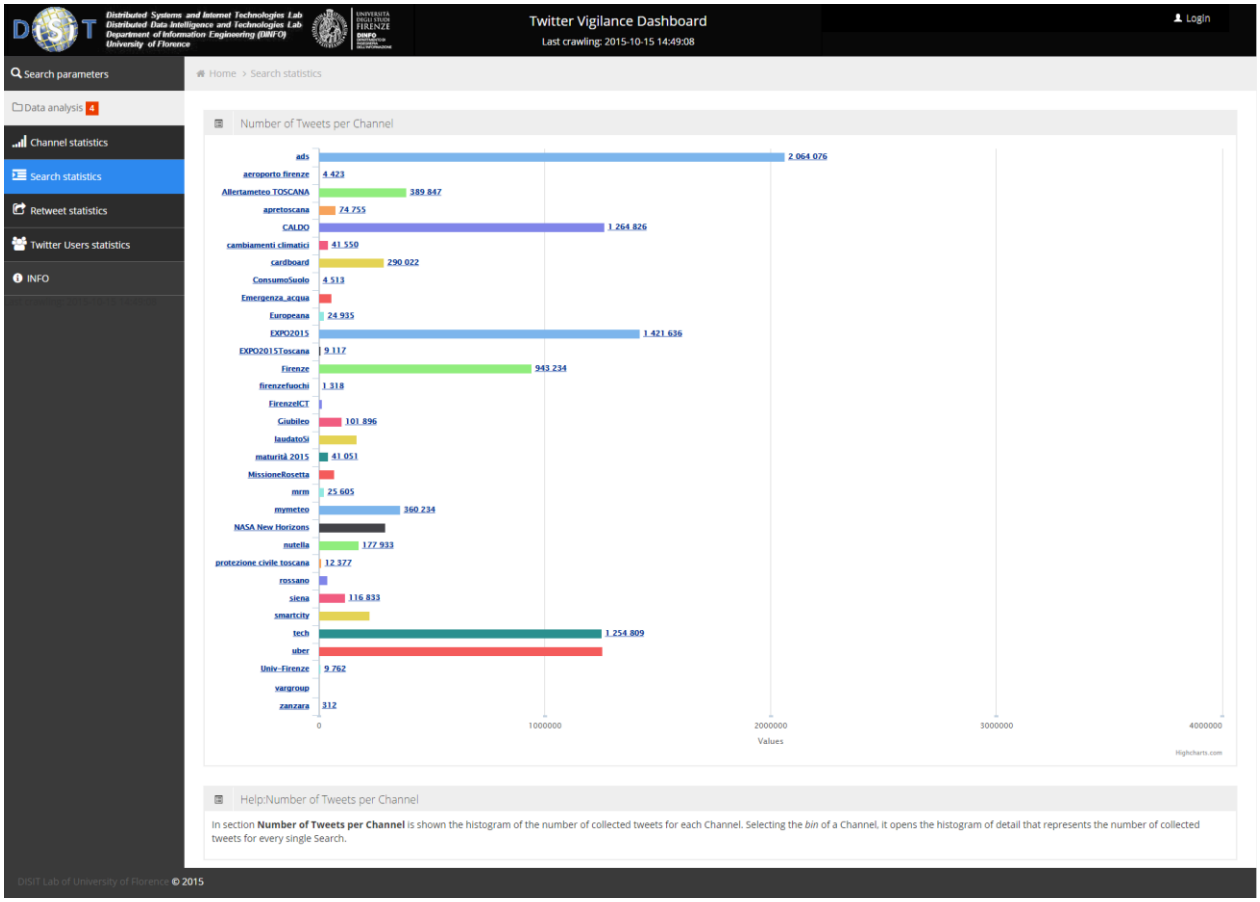


Figure 4: Search statistics page

Clicking on a bin it is possible display another histogram that shows the total values of the messages of the individual searches that are part of channel, as shown in Figure5.

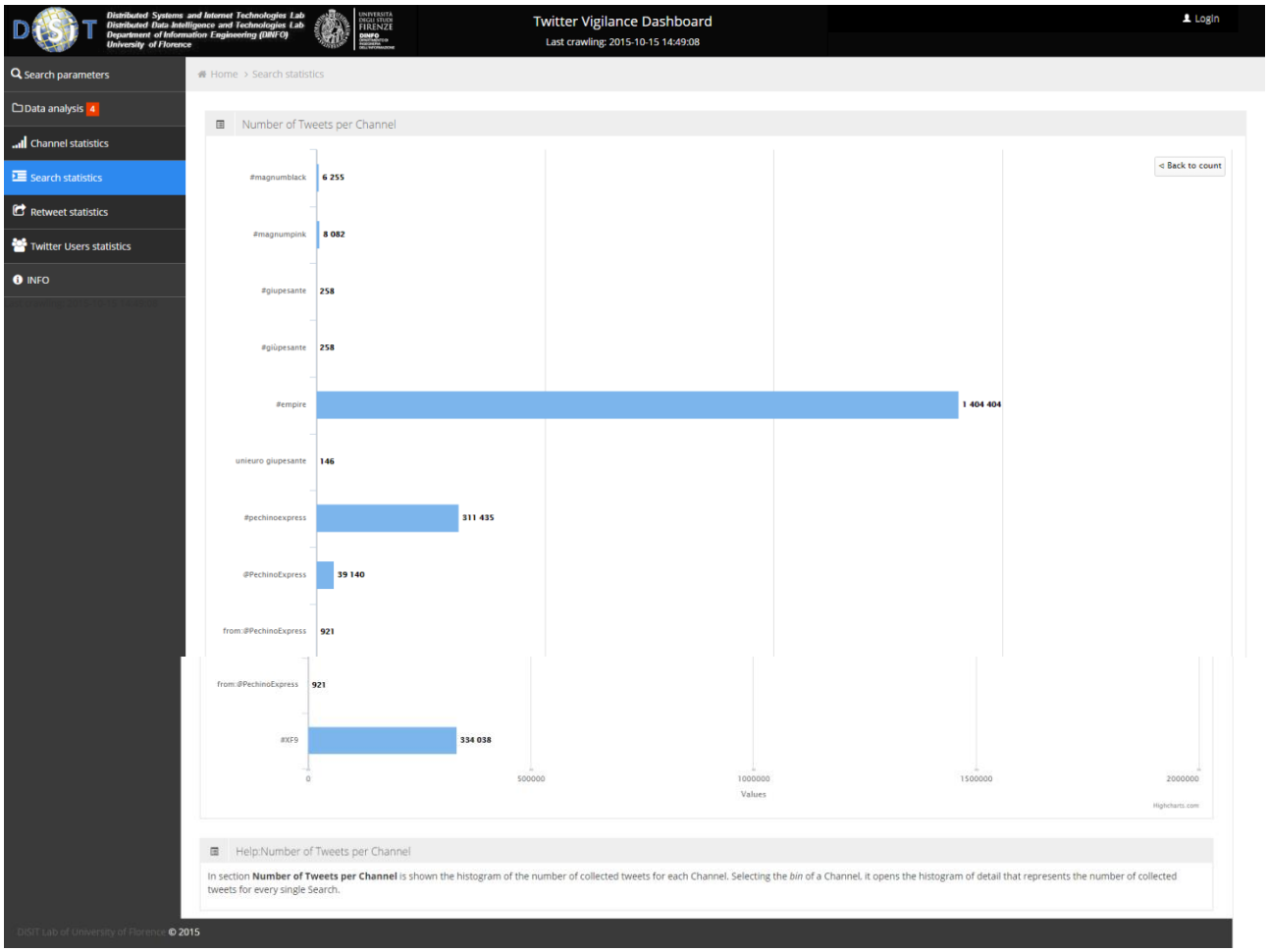


Figure 5: Search statistics page: single channel details

2.2 Twitter User Statistics

In this page (Figure6) there are a histogram and a table representing the same data, i.e. the total number of distinct users for each channel.

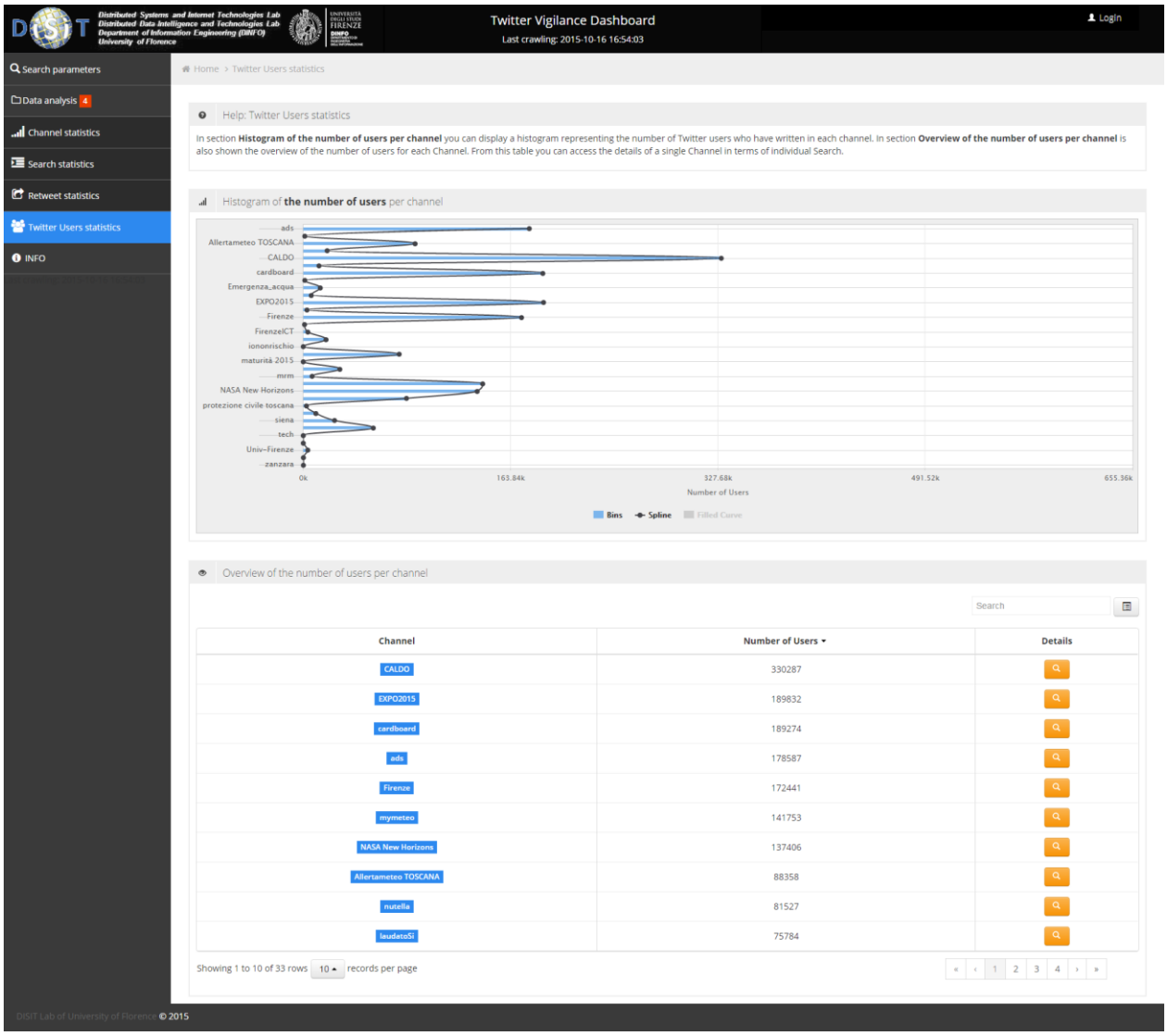


Figure 6: Twitter Users statistics

Clicking on the button in the Details column of the table leads to the detail page of the channel, Figure, where there are two graphs and a table: The table represents the number of messages related to individual searches for that channel, the histogram shows the top 10 users with the greatest number of posts on the channel and the pie chart shows the distribution of users in individual research belonging to the channel.

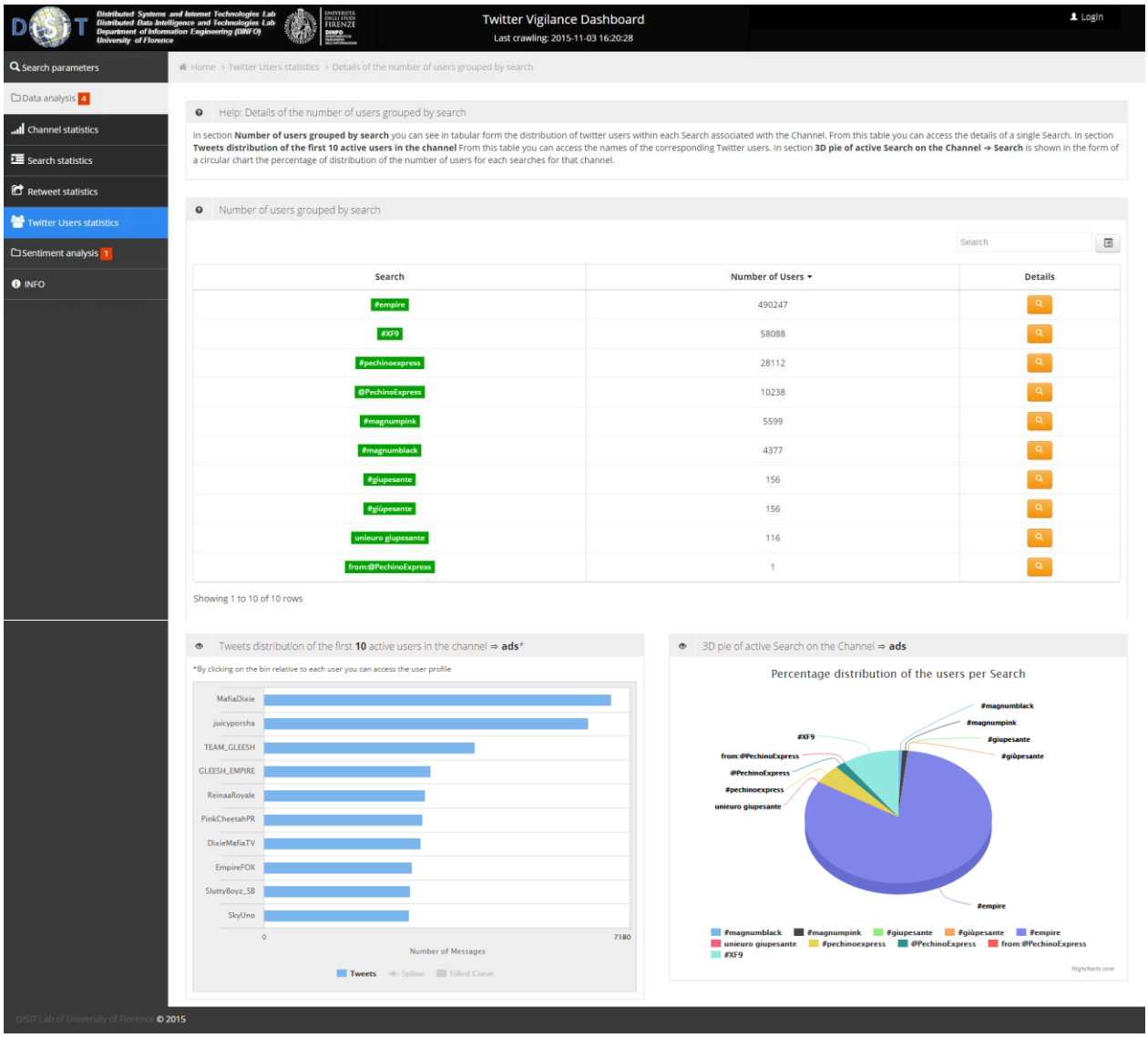


Figure 7: Details of the number of users grouped by search

Selecting the bin of a user it's possible to access the profile of the user.

Clicking on the button in the Details column of the table leads to the detail page of the Search, Figure, where is displayed a table: the table represents the list of users who have written at least one message among those recovered from the Search and the number of posts written by each user.

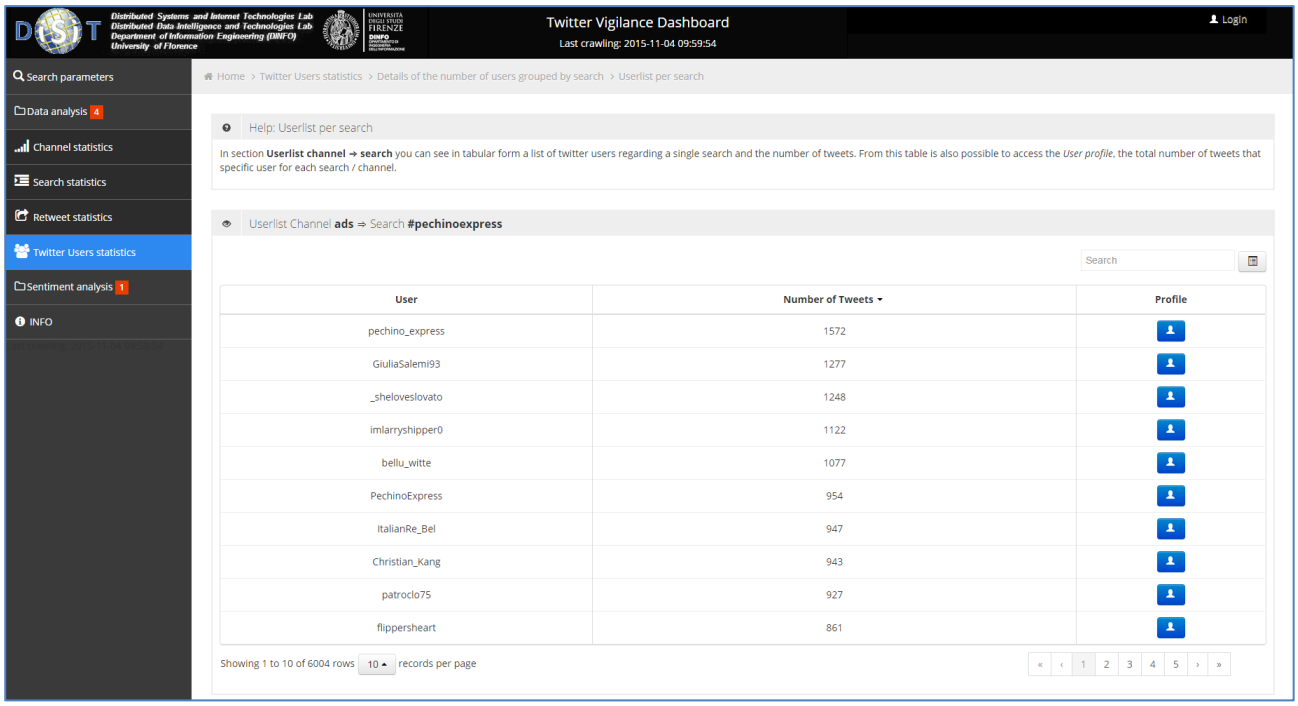


Figure 8: Details of user grouped by Search

Clicking on Profile button it's possible to access the profile of the user (Figure).

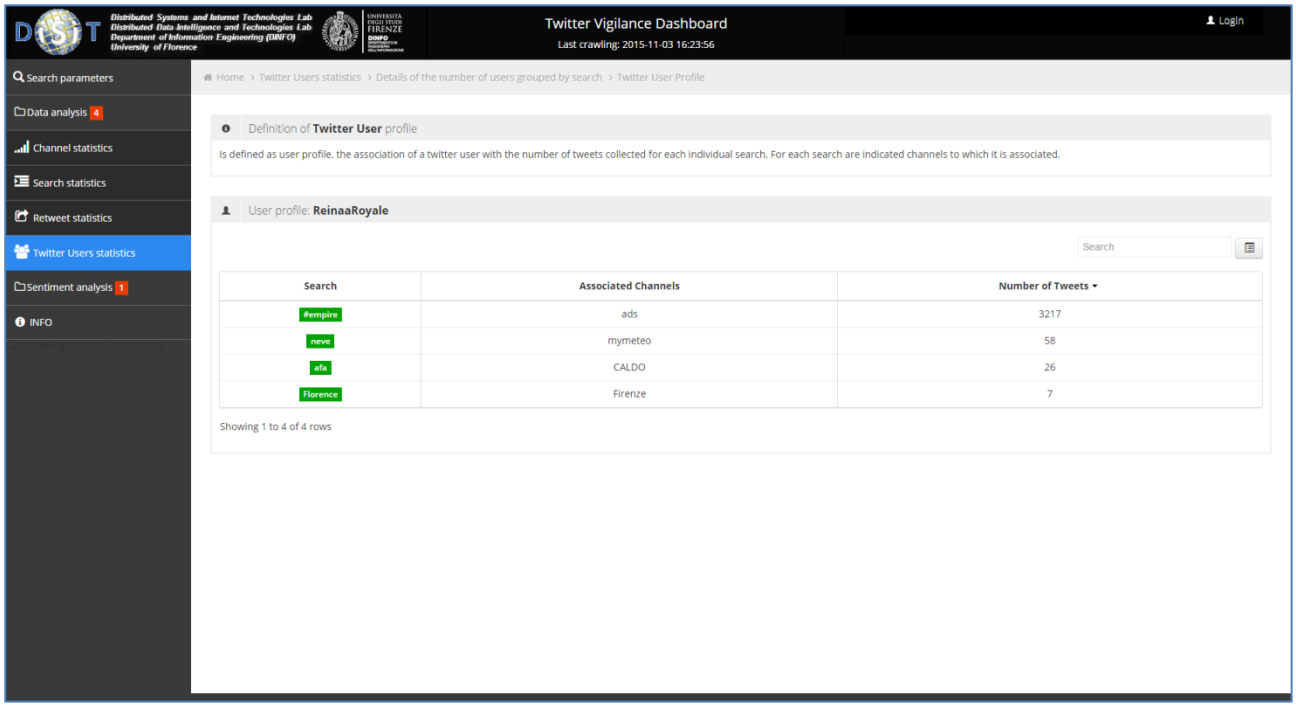


Figure 9: User profile

The Figure displays the association of a twitter user with the number of tweets collected for each individual search. For each search outlines the channels to which it is associated.

2.2.1.1 Retweets

This page displays a bar chart that shows the number of tweets and retweets for each channel.

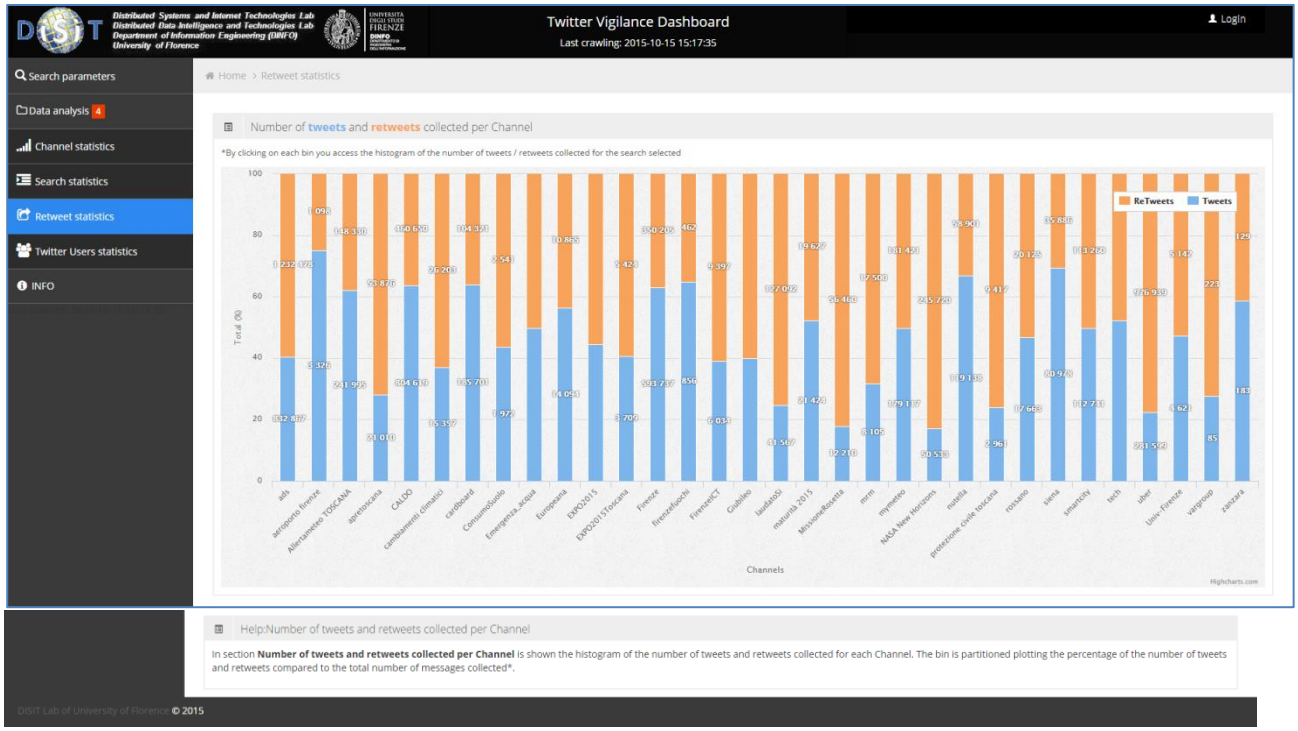


Figure 10: Retweet statistics page

Clicking on a single bar is shown the details of the research associated with the selected channel

Figure 11: Retweet statistics for single Channel page

3 Modelli di Information Extraction da dati e contributi testuali acquisiti tramite Social Media e Crowd Sourcing

Following two case studio:

- Transportp Publico Locale (TPL);
- Uber.

Figura 1: Global View

3.1 TPL, Trasporto Pubblico Locale

Working steps:

- Channel setup
- Noise filtering
- Customer satisfcation analysis among different operators
- Customer compliant classification

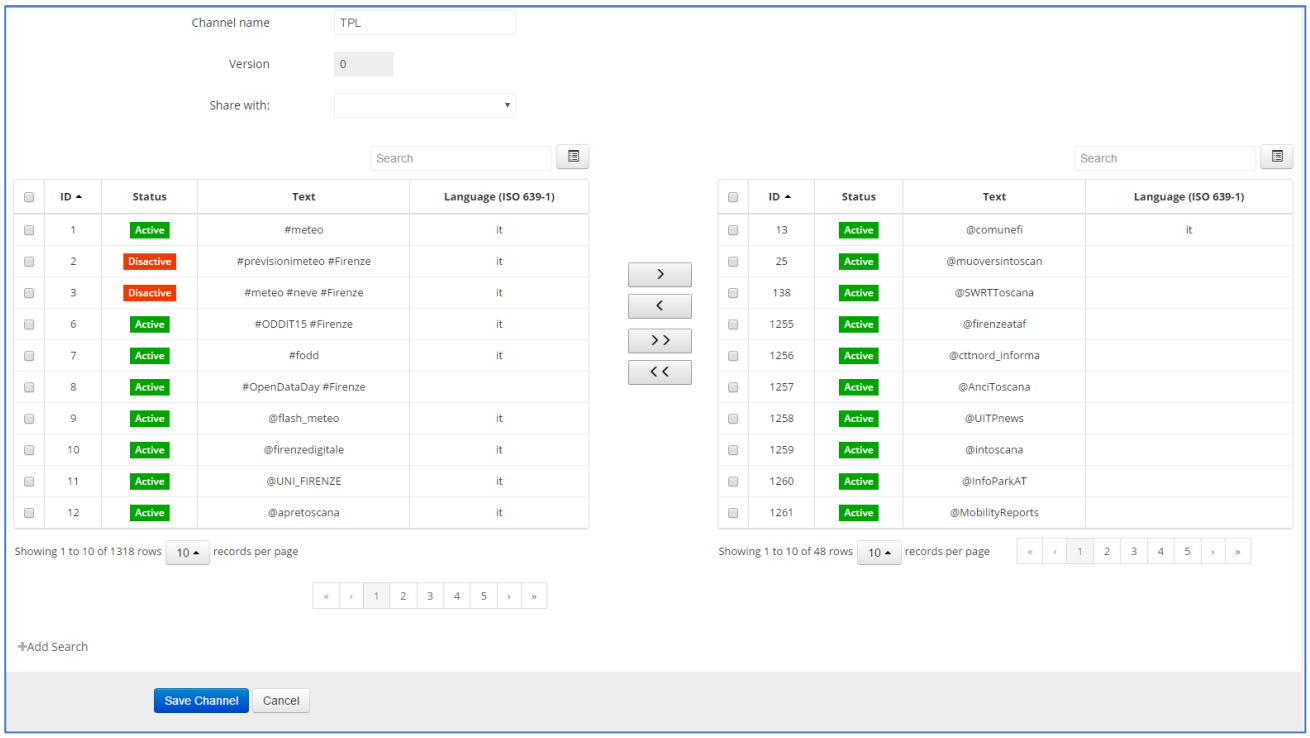


Figura 2: TPL channel setup

The TPL channel is composed of the following searches: #bus #fipili #intreno #publictransport #tramviafi #travel #trenitalia @AMTToscana @AnciToscana @ArezzoPendolari @AutolineeCurcio @AutolineeRomano @CAPautolinee @capviaggiprato @cispeltos @Clickmobility @comunefi @CTM_Cagliari @cttnord_informa @esserependolare @EuroTransMag @ferpress @firenzeataf @GroupeRATP @iMobChallenge @iMobilityForum @InfoBusPisa @InfoParkAT @intoscana @ItaloTreno @LAMIAFERMATA @LeFrecce @MobilityPress @MobilityReports @muoversintoscan @OrariBus @OssMobProvPI @pendolarifr2 @PiuBus @StazioniSicure @SWRTToscana @tolcommunity @ToremarFerries @Toscanaeturismo @tranviafirenze @TrasportiItalia @TTSItalia @UITPnews

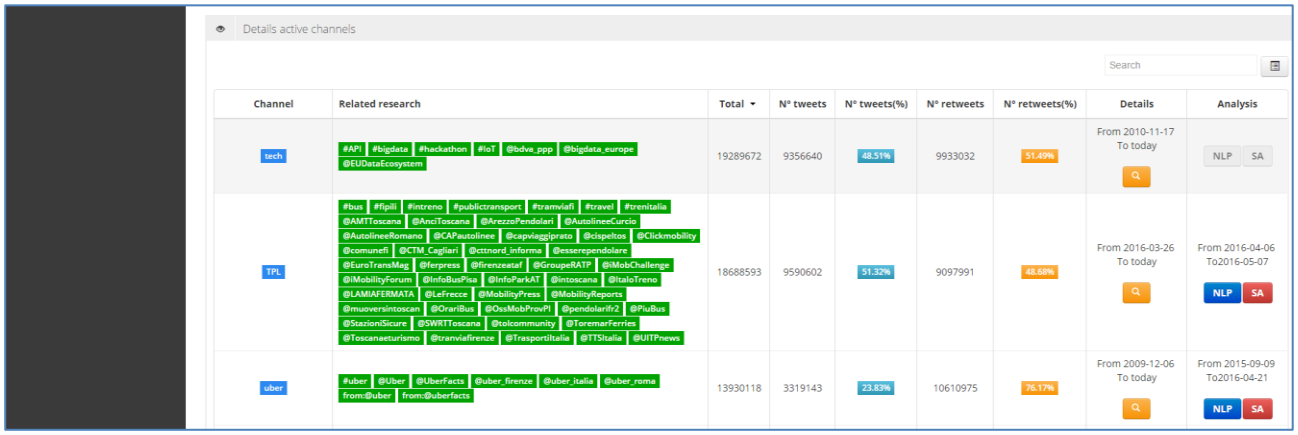


Figura 3 TPL detailed stats

Figure 3 shows a total of 18732579 Tweets which 9617347 are tweets while 9115232 are retweets.

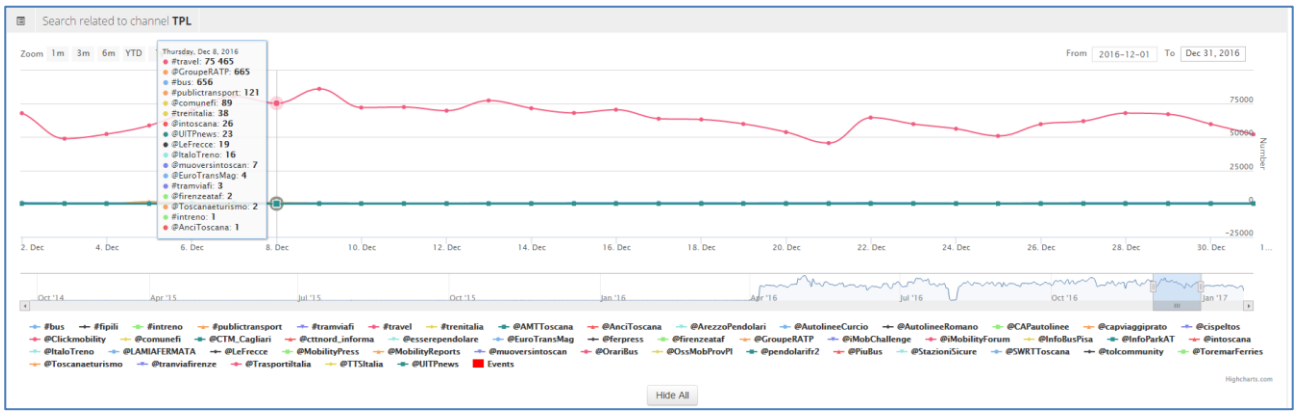


Figura 4 Search related to channel TPL

Figure 4 shows that #travel (pink line) is dominant (10 times magnitude compared with others)

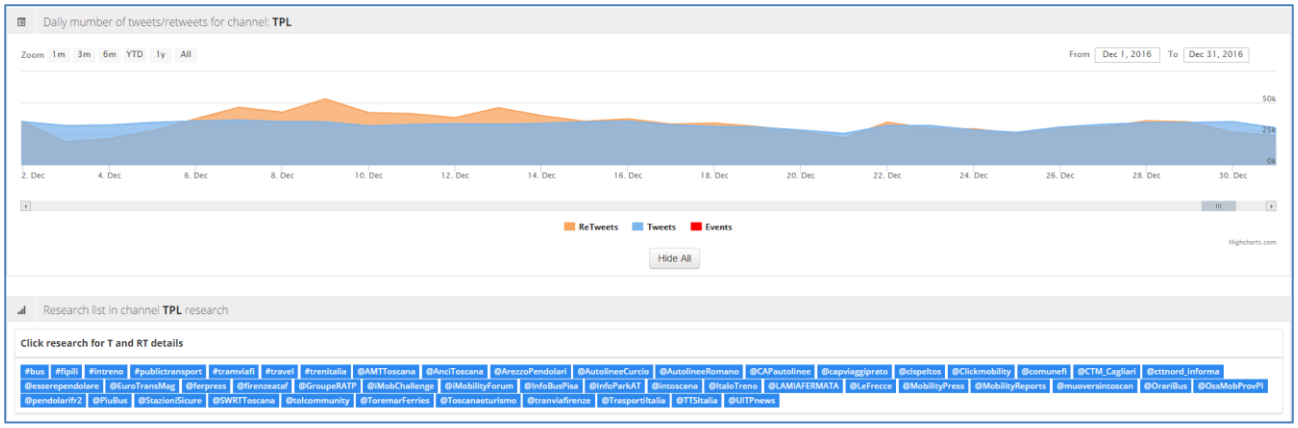


Figure 5 Daily number of tweets/Retweets for channel TPL

Figure 5 shows abnomral activity in the range starting with 6/12 and ending with 14/12 which highly correlated with national public trasportation strike of 9/12 [1].

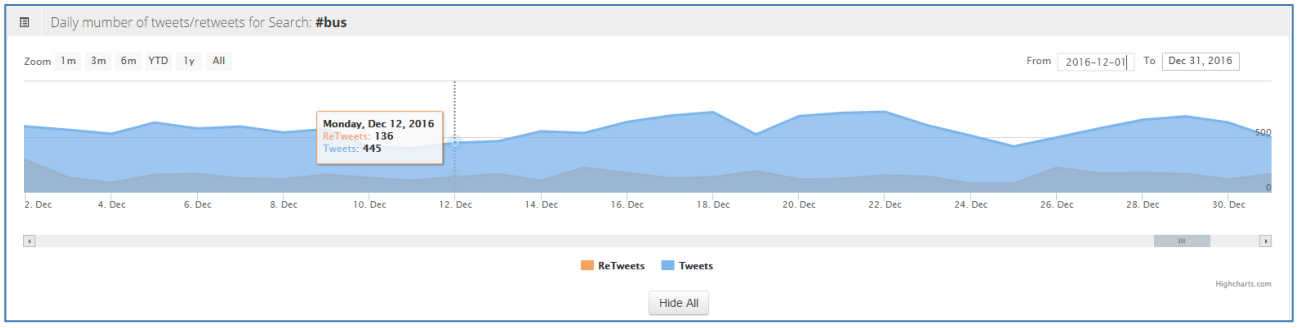


Figure 6 Daily number of tweets/Retweets for single search

3.1.1 NLP

The NLP analysis is highly influenced by tweets belonging to #travel. Figure 7, 8, 9 and 10 shows dominant English terms (e.g. #must, #extreme, @airbnb, @americanair etc ..)

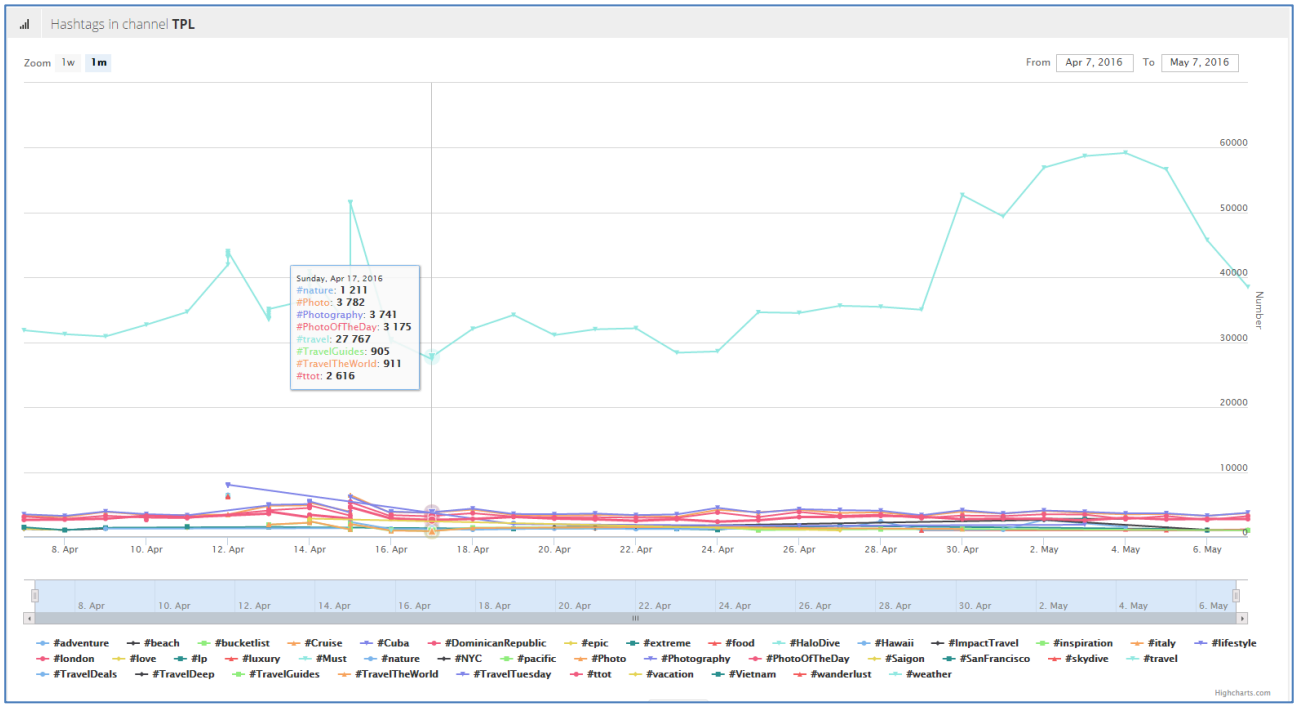


Figura 7 Hashtags for TPL

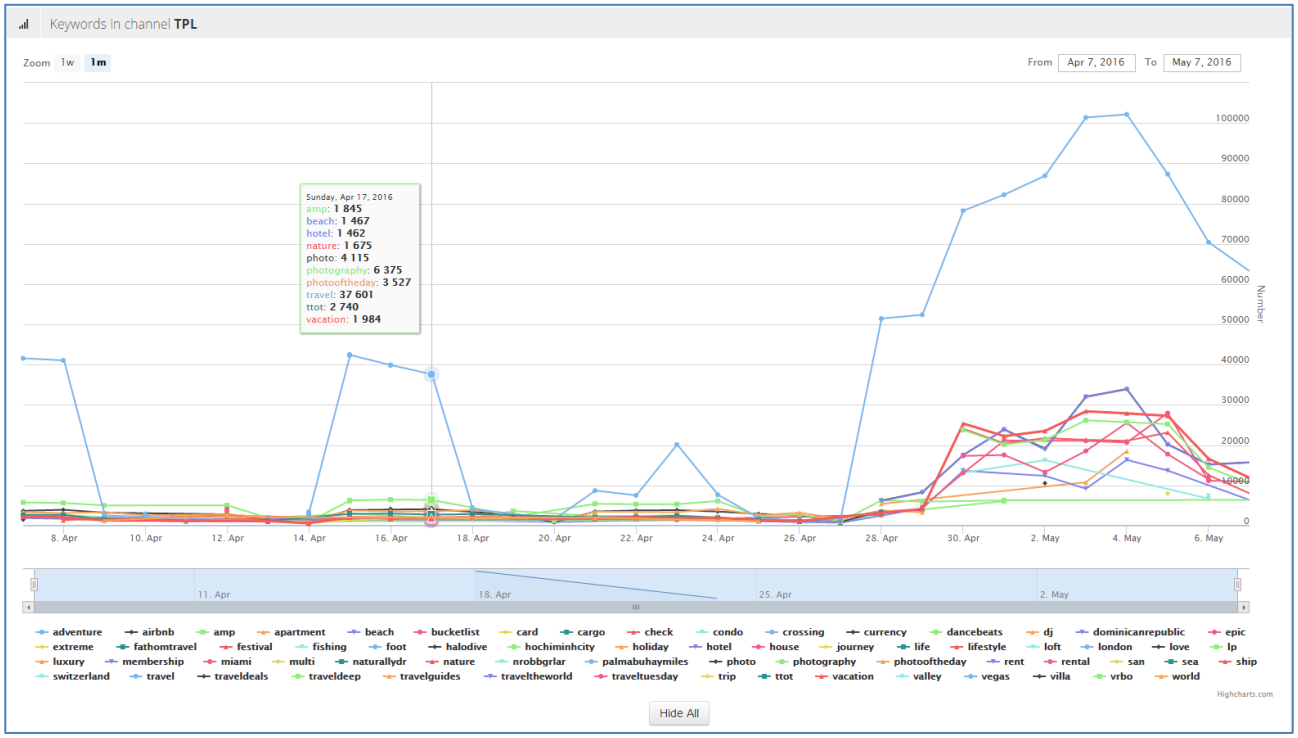


Figura 8 Keywords for TPL

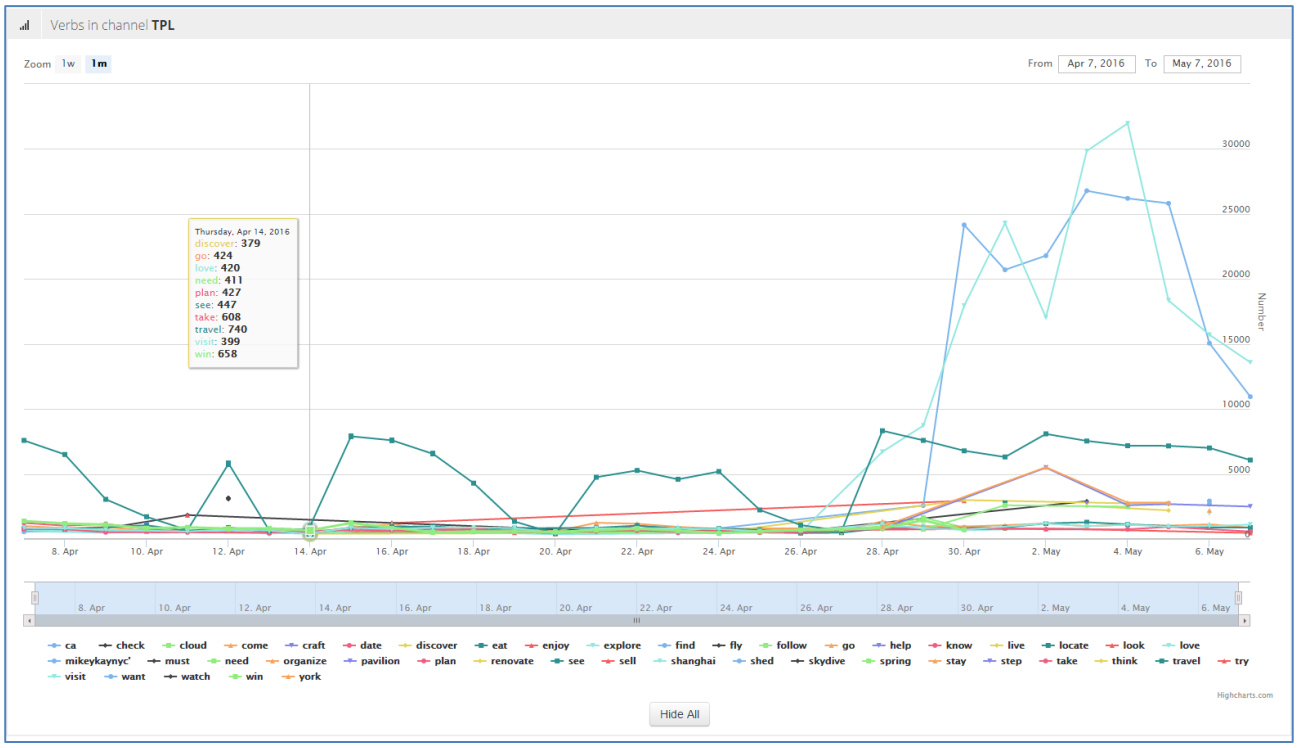


Figura 9 Verbs for TPL

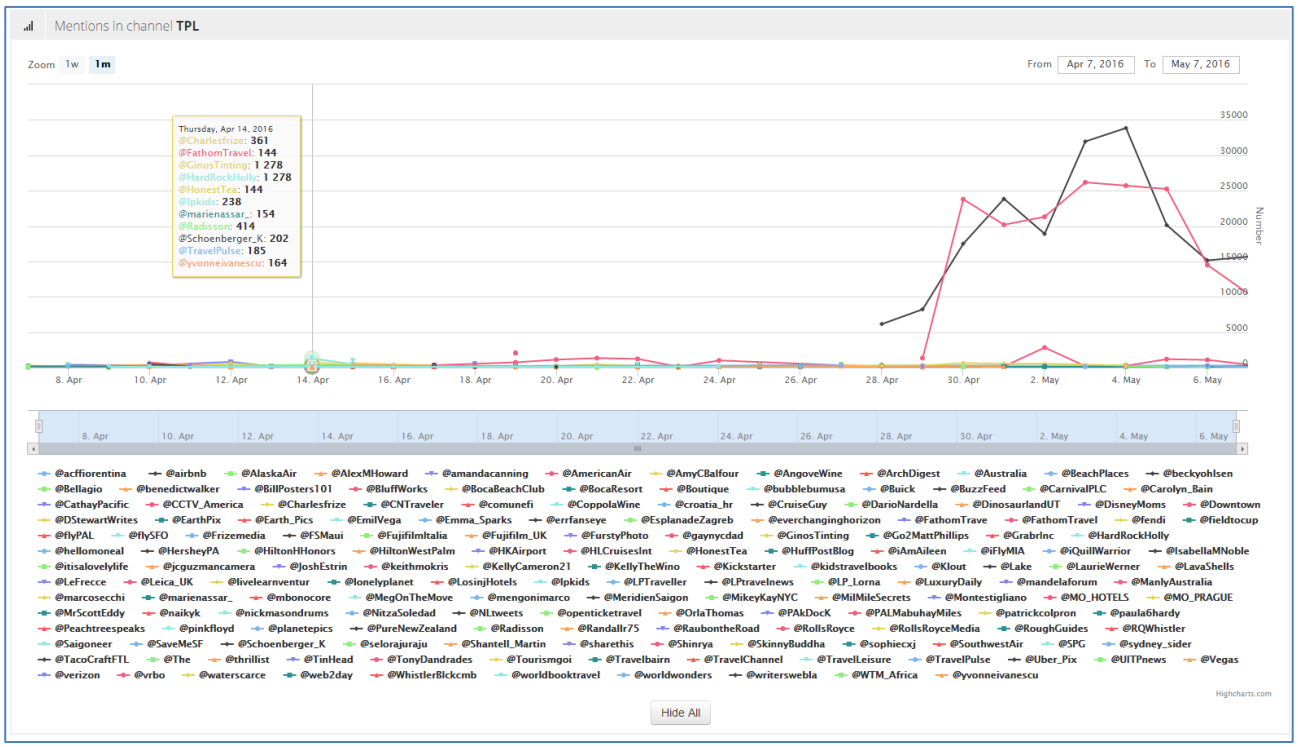


Figura 10 Mentions for TPL

3.1.2 Sentiment

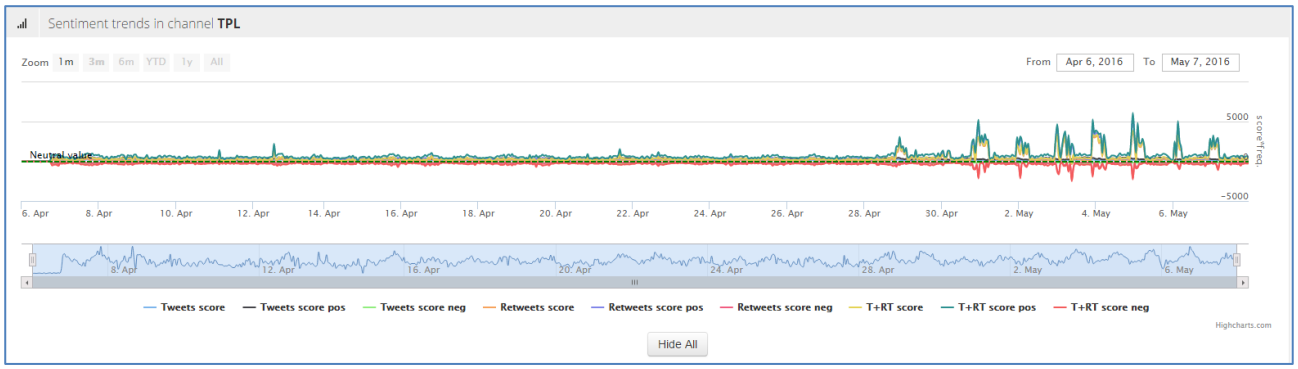


Figura 11 Sentiment trends in TPL

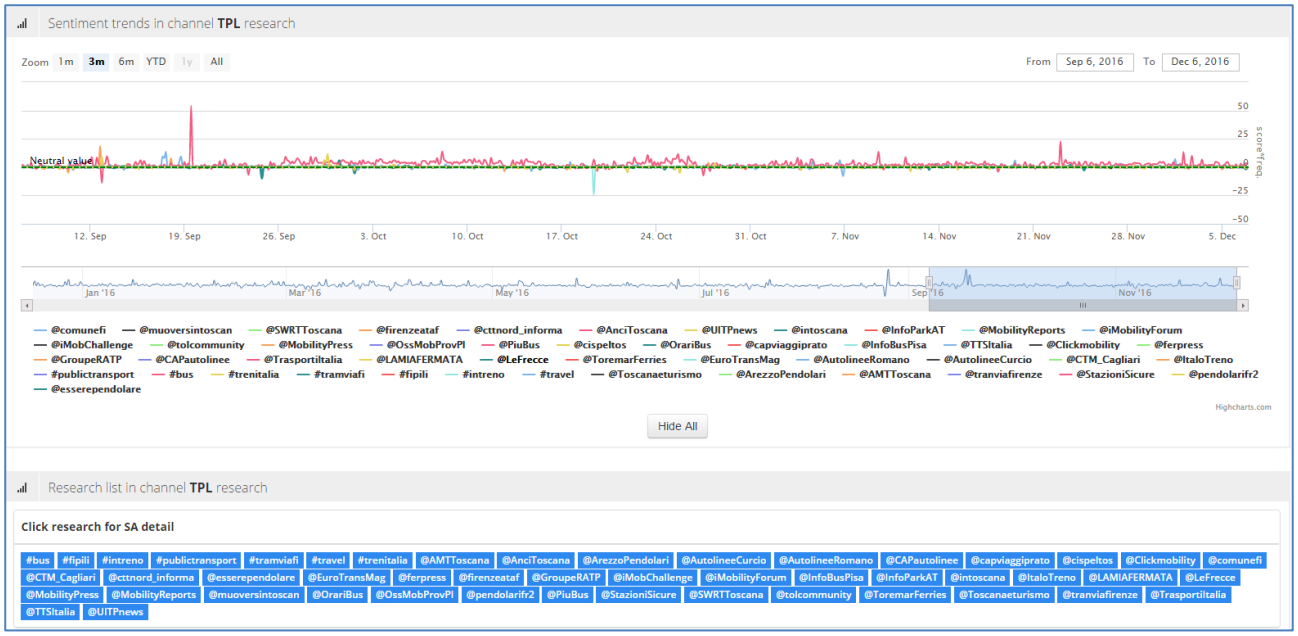


Figura 12 Sentiment trends in TPL researches

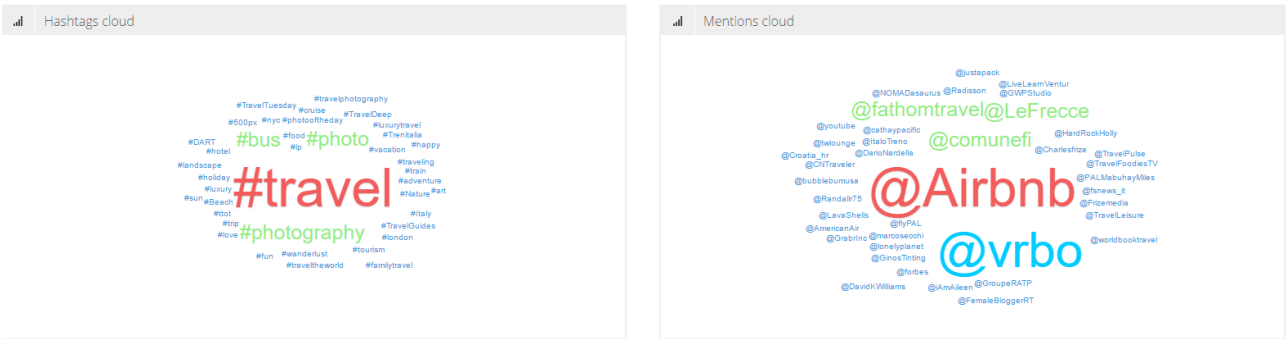


Figura 13 Word cloud

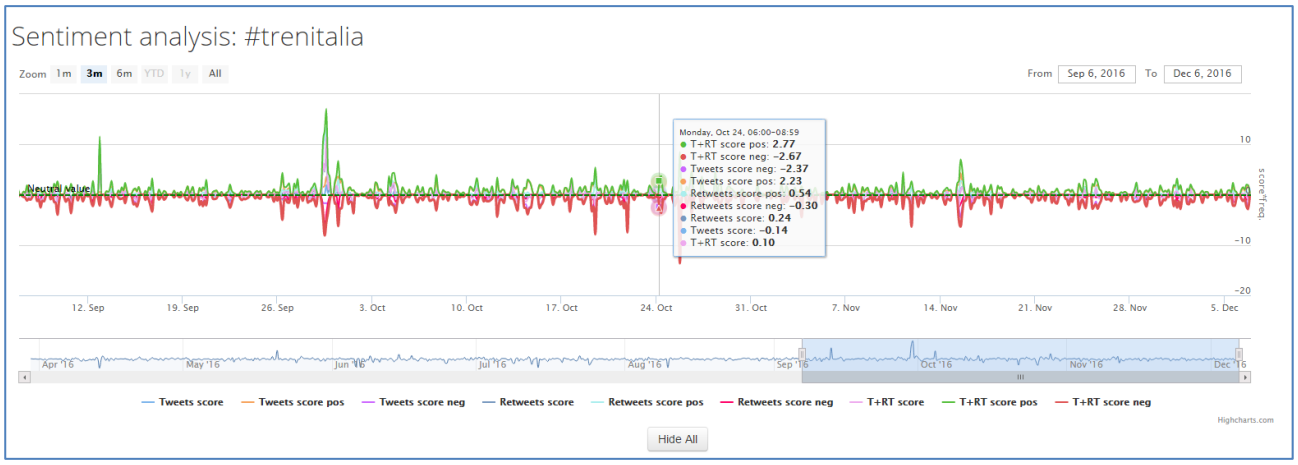


Figure 14 Sentiment analysis per single search

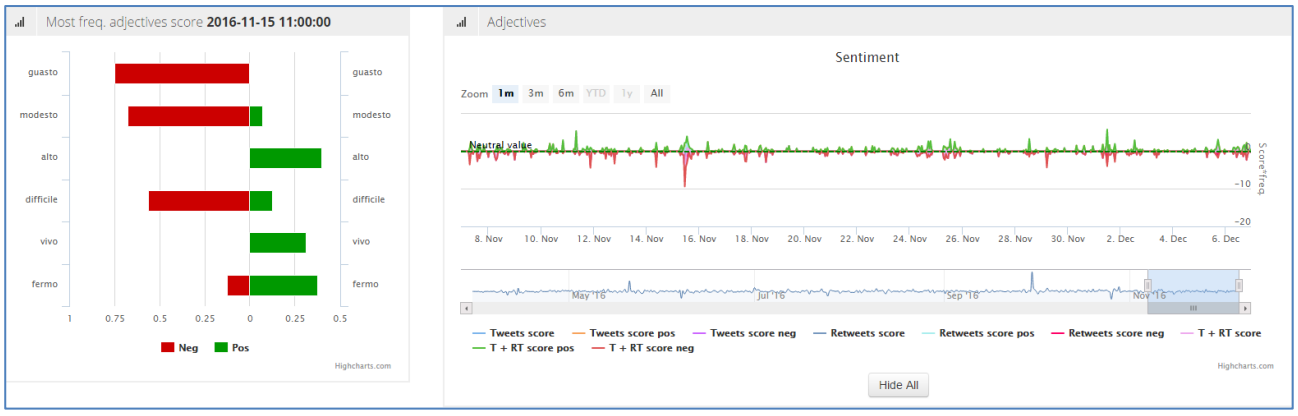


Figure 15 Sentiment per adjective with most frequencies per single search

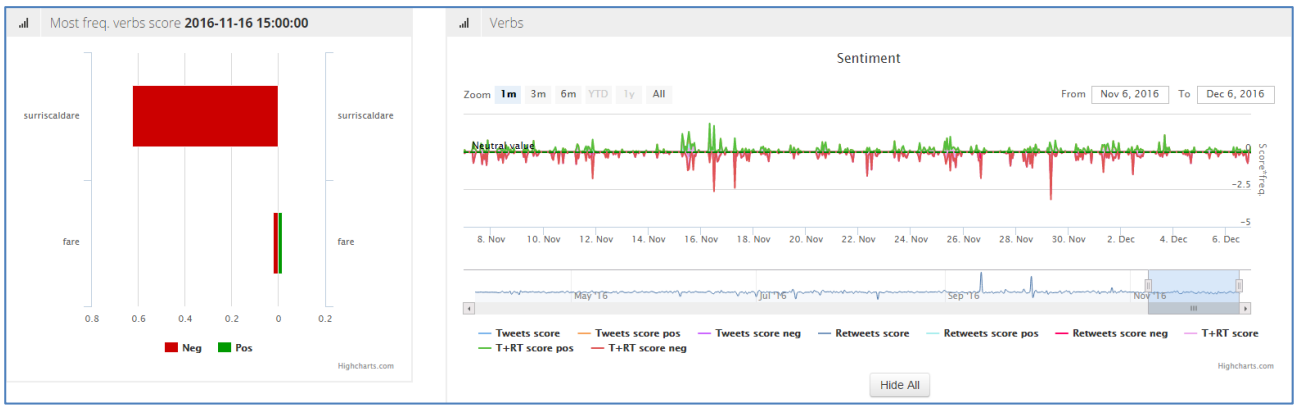


Figure 16 Sentiment per verbs with most frequencies per single search

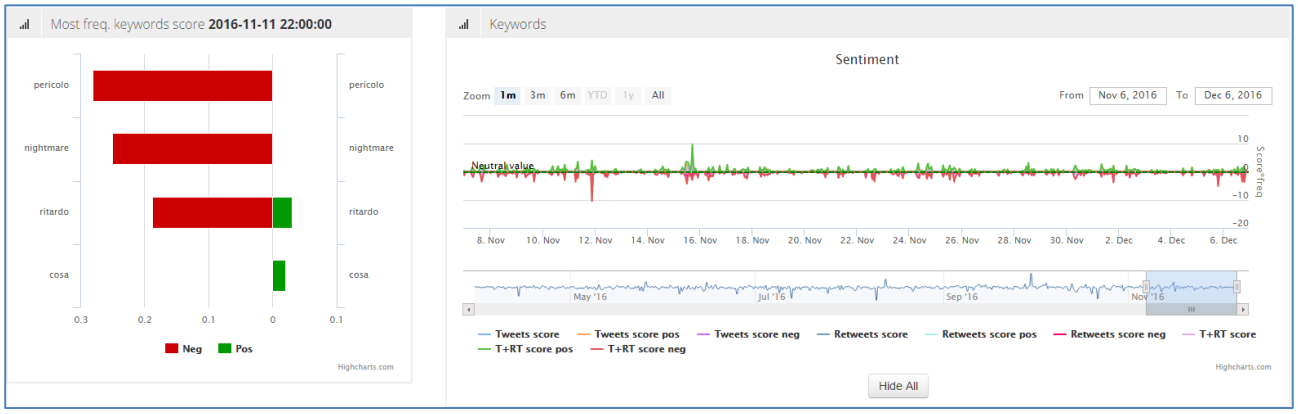


Figure 17: Sentiment per verbs with most frequencies per single search

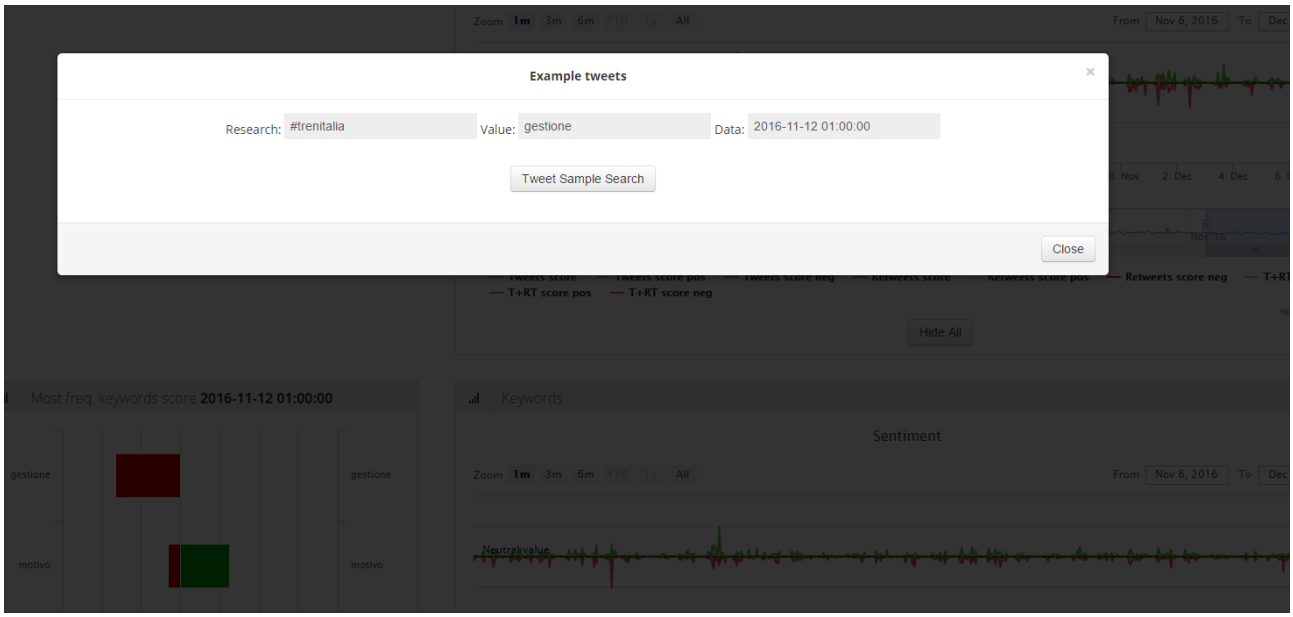


Figura 18 Advanced tweets search

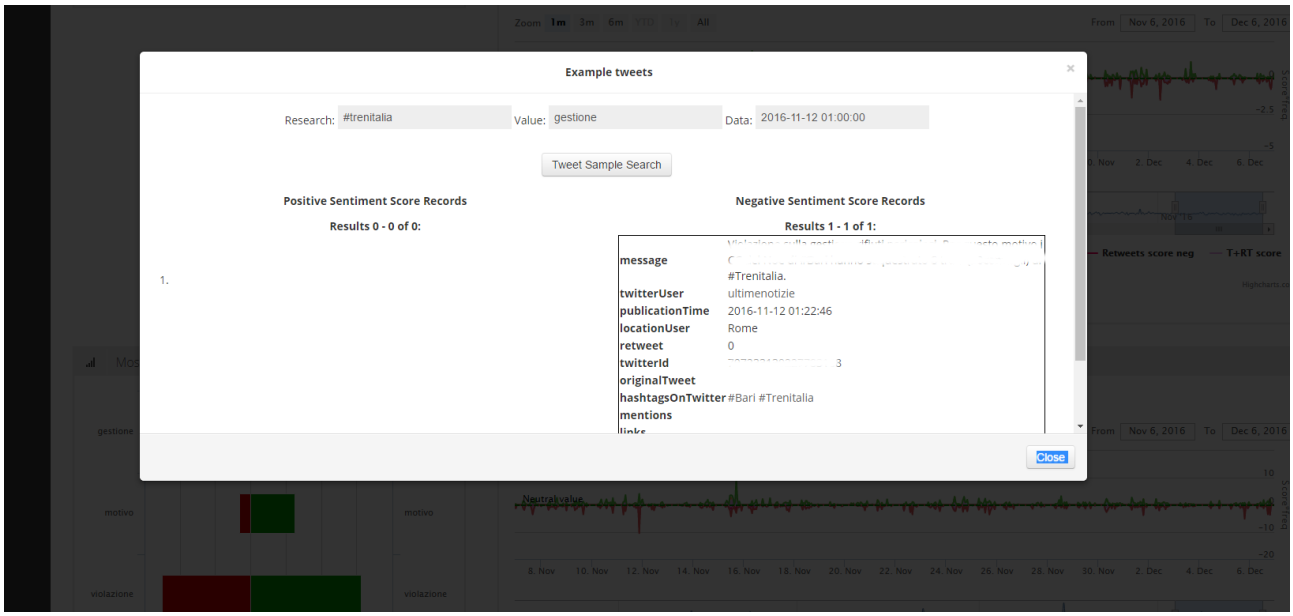


Figura 19 Example result

3.2 UBER

Working steps:

- Channel setup
- Noise filtering
- Customer satisfcation analysis
- Customer compliant classification

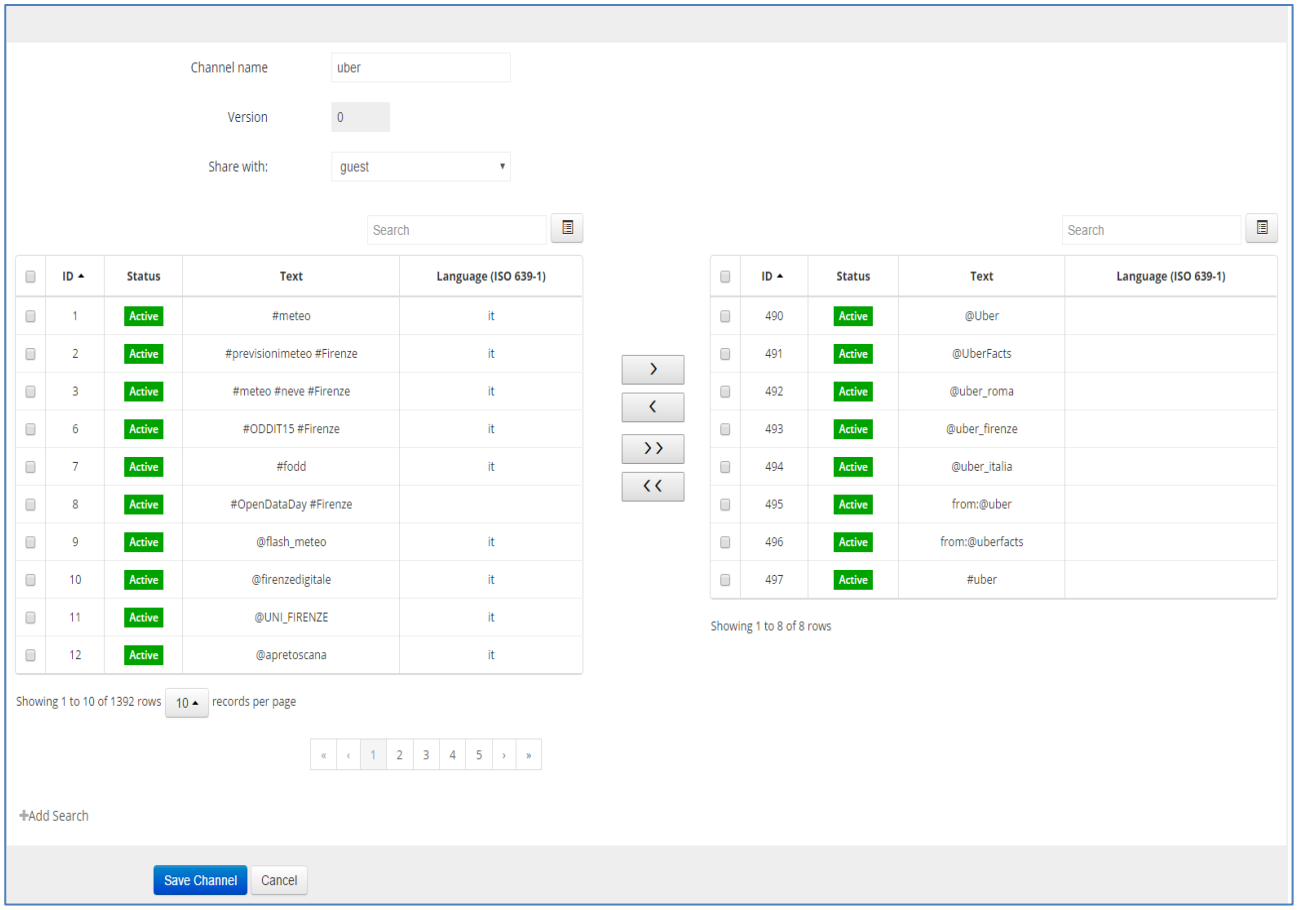


Figura 20: UBER channel setup

The UBER channel is composed of #uber @Uber @UberFacts @uber_firenze @uber_italia @uber_roma from:@uber from:@uberfacts.

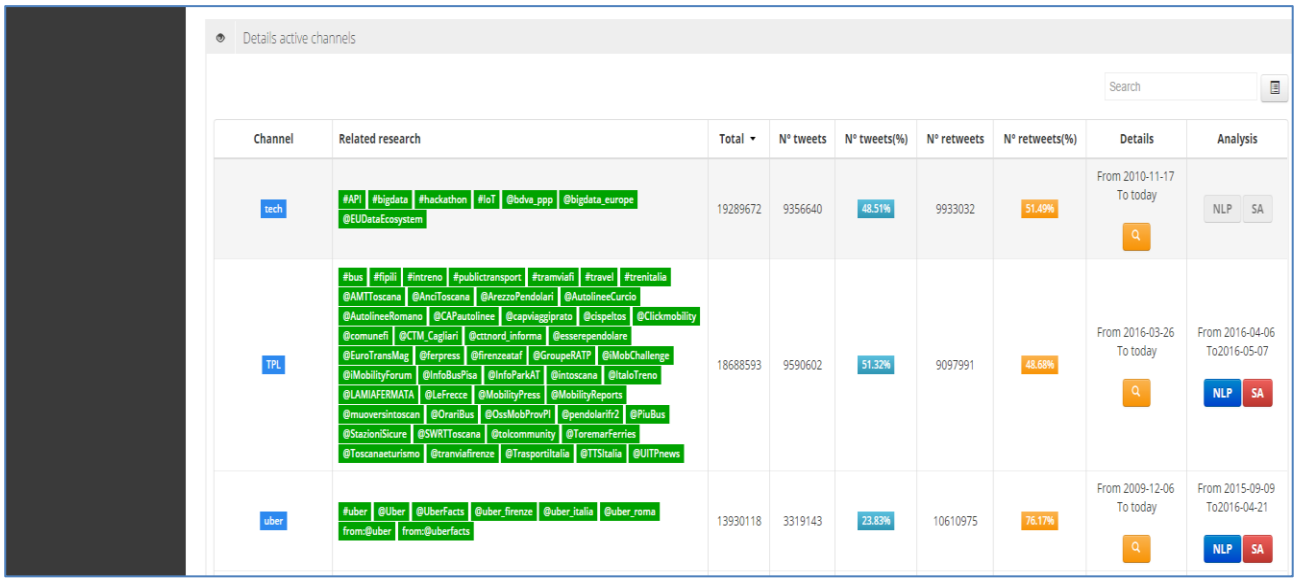


Figure 21: UBER detailed stats

Figure 21 shows a total of 13944352 Tweets which 3324272 are tweets while 10620080 are retweets.

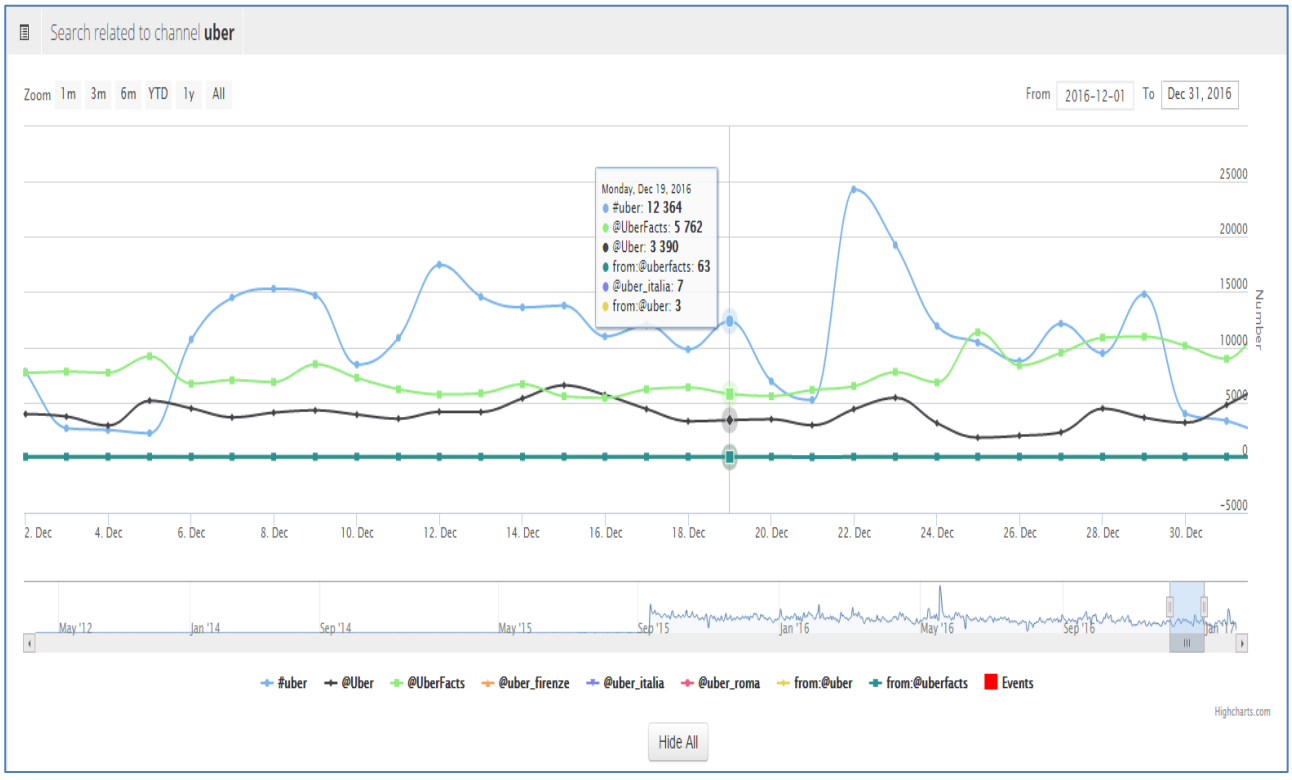


Figure 22: Search related to channel UBER

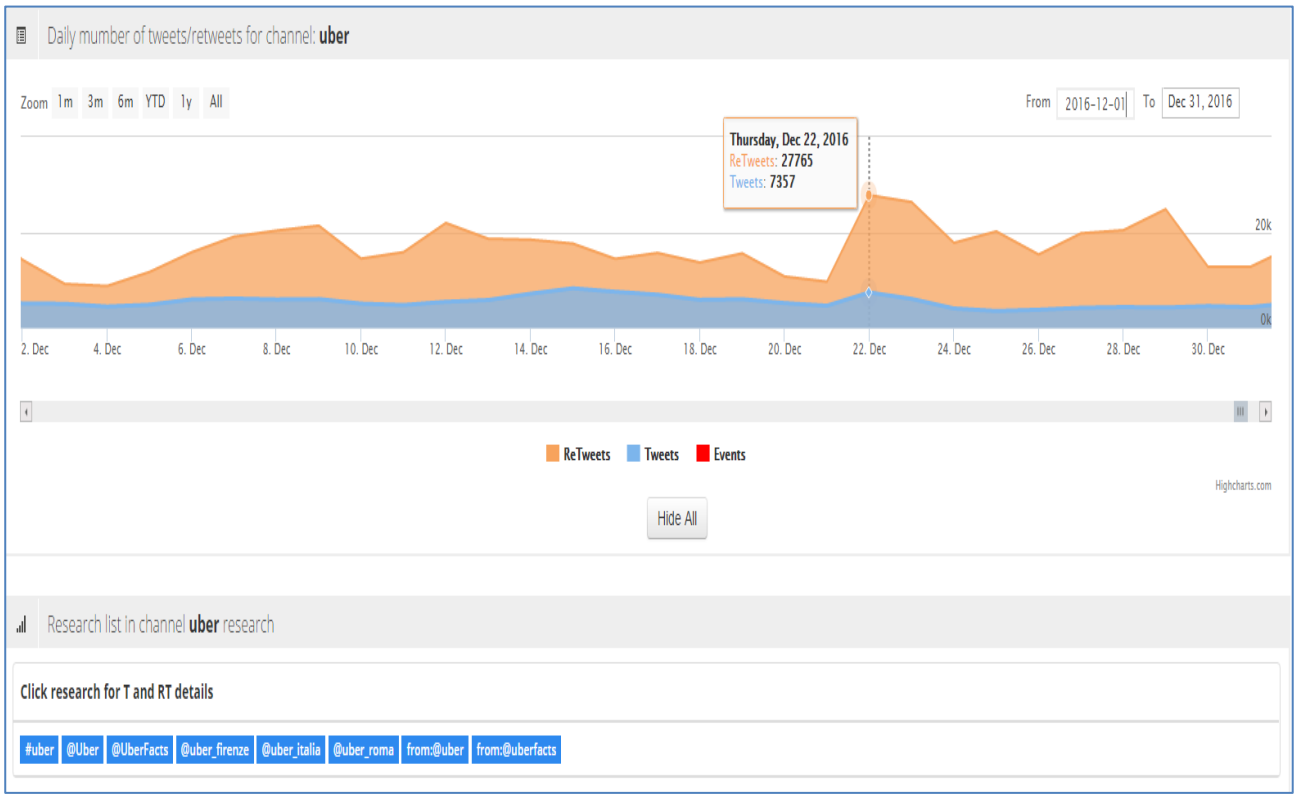


Figure 23: Daily number of tweets/retweets for channel UBER

3.2.1 NLP

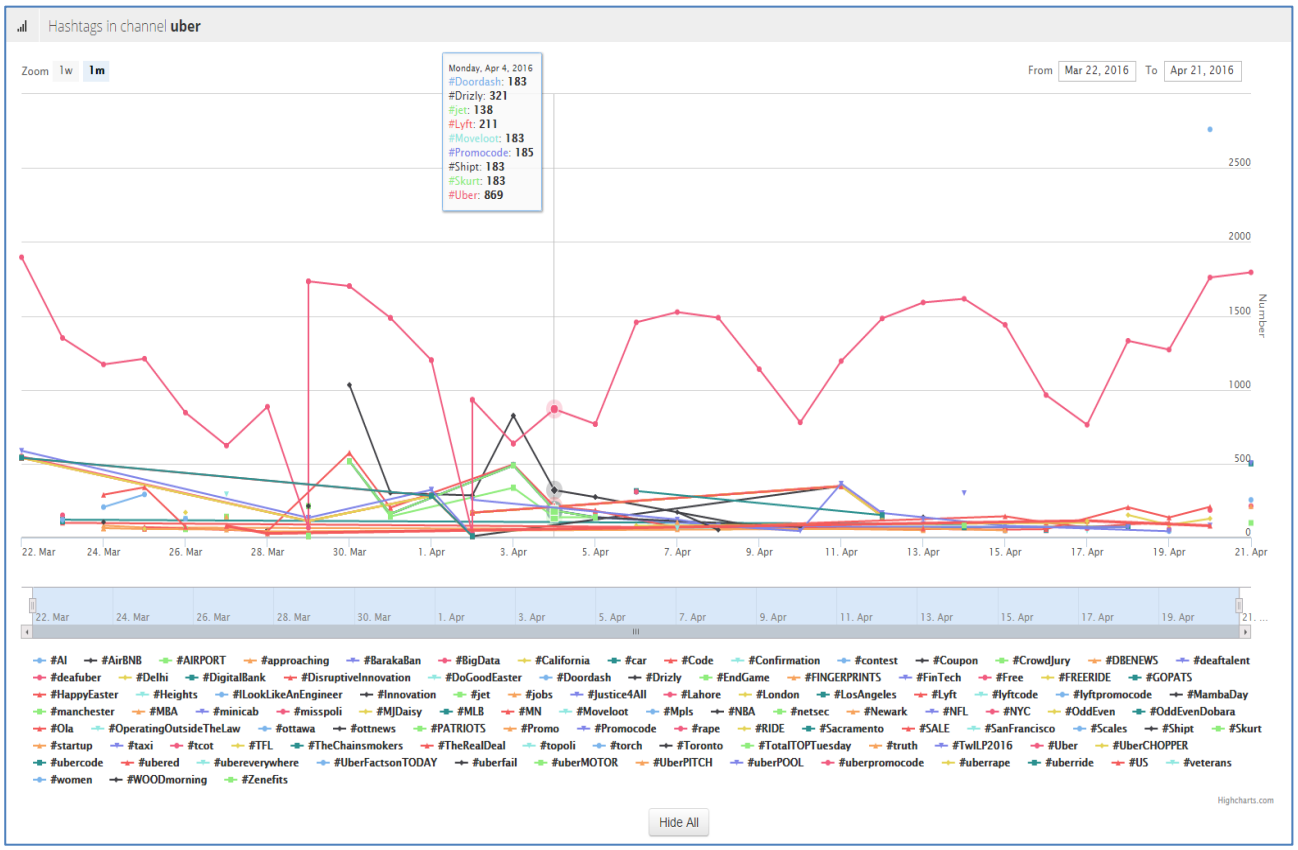


Figure 24: Hashtags for UBER

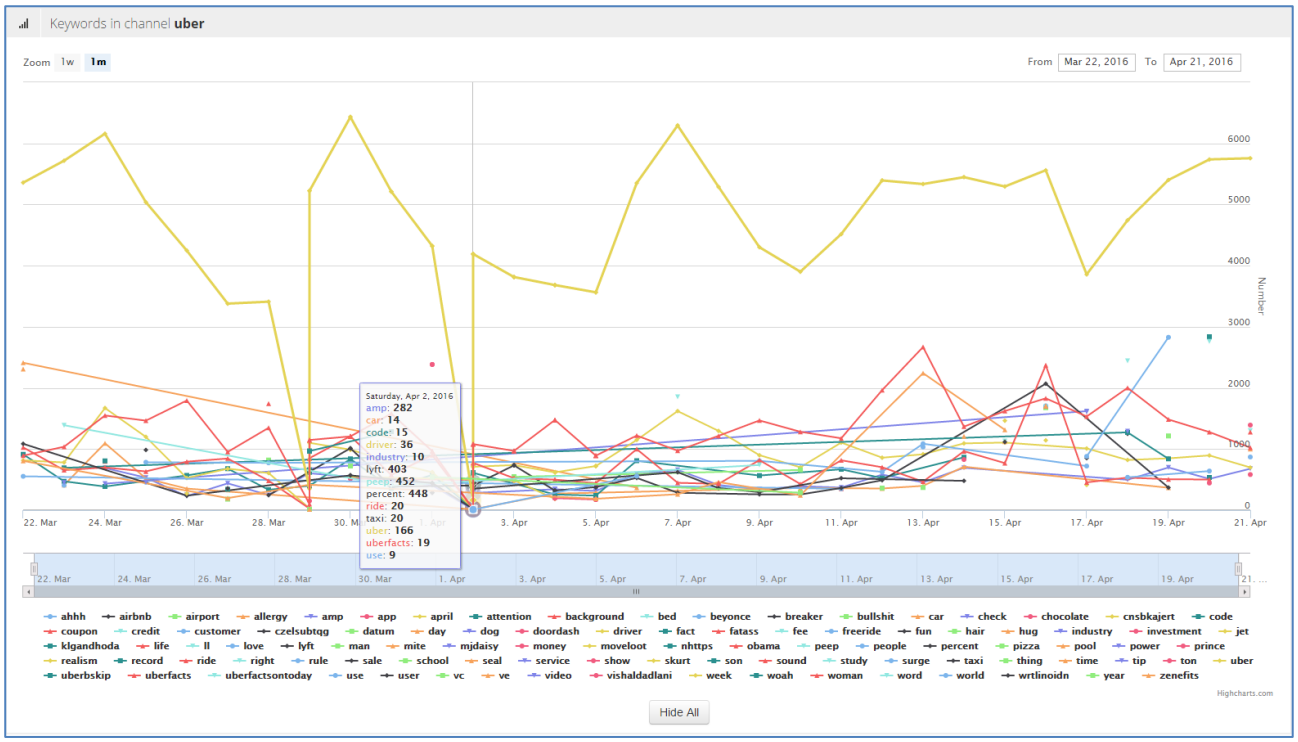


Figura 25: Keyword for UBER

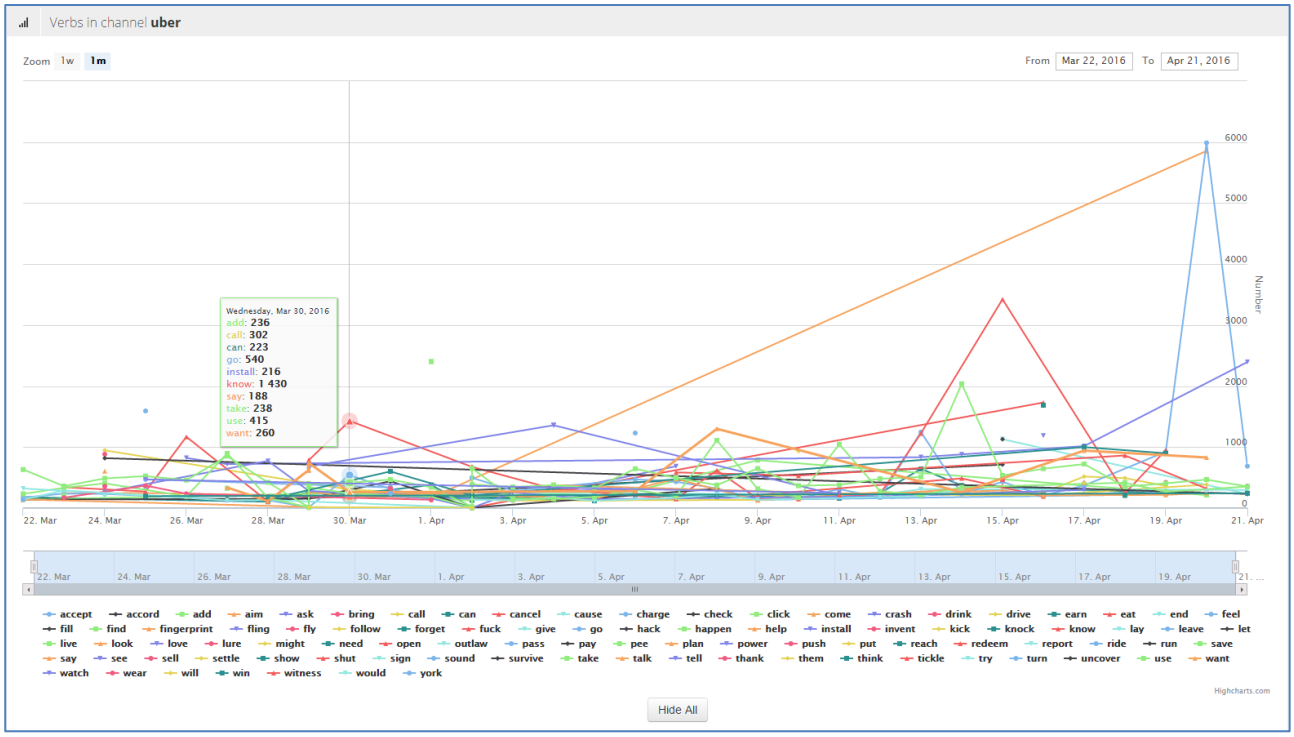


Figura 26: Verbs for UBER

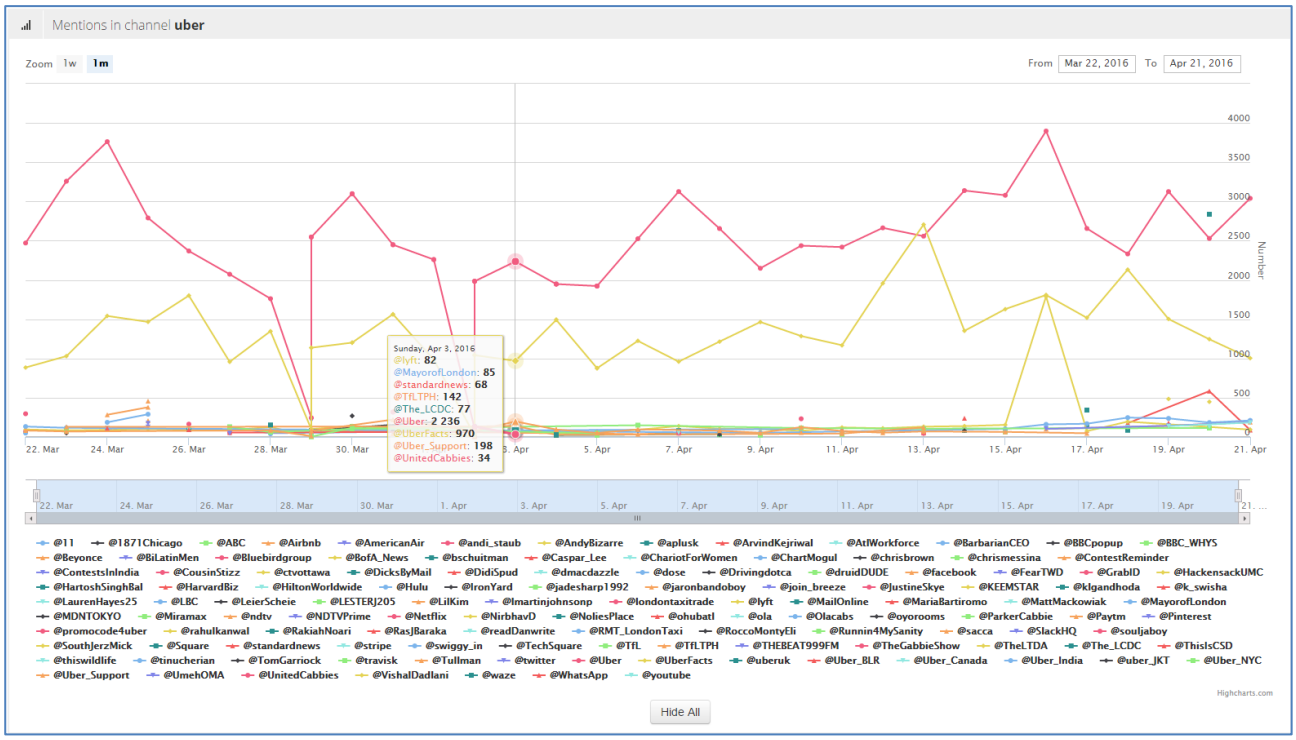


Figura 27: Mentions for UBER

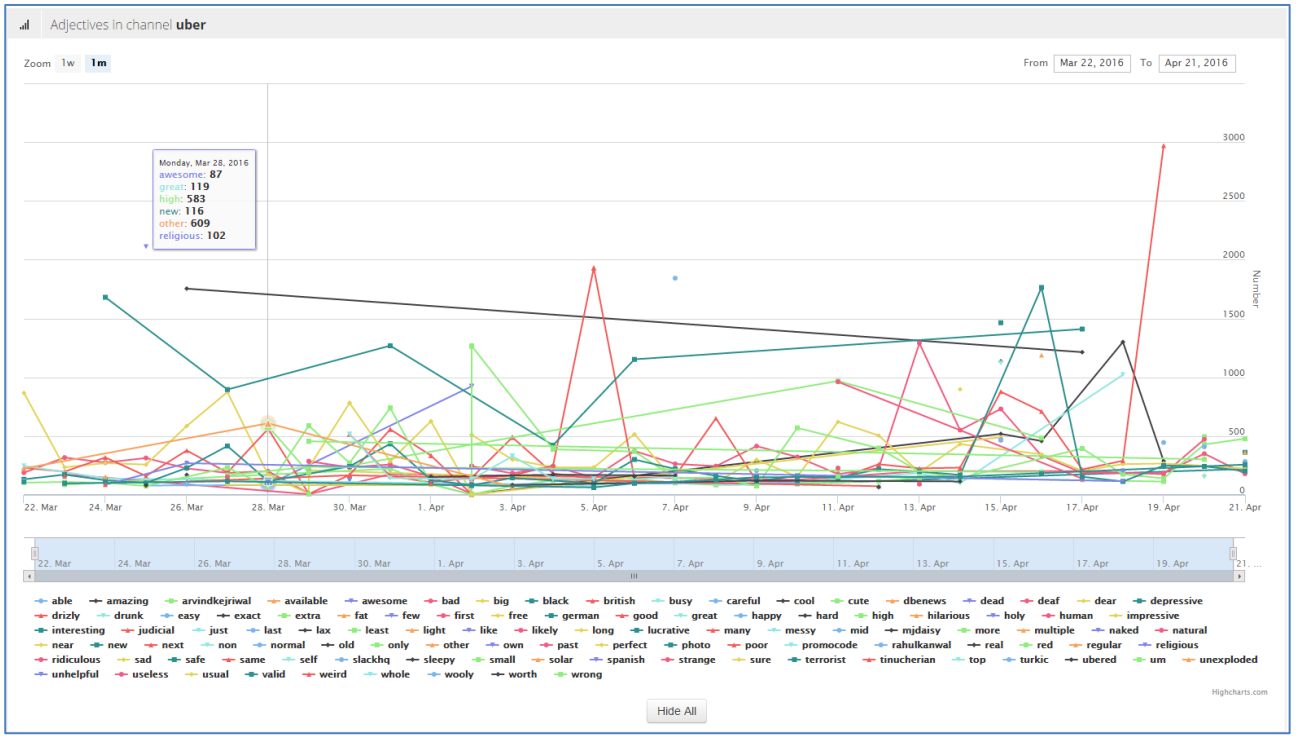


Figura 28: Adjective for UBER

3.2.2 Sentiment Analysis

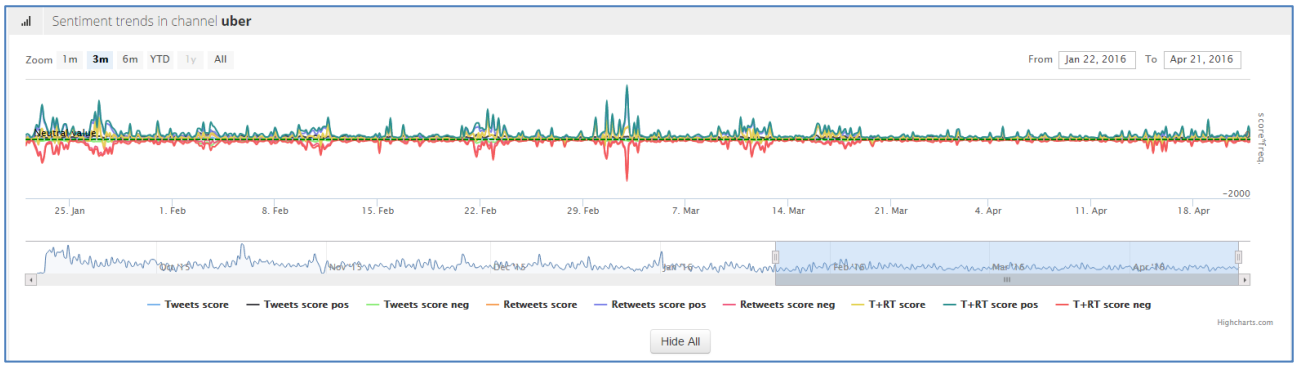


Figura 29: Sentiments trends in UBER

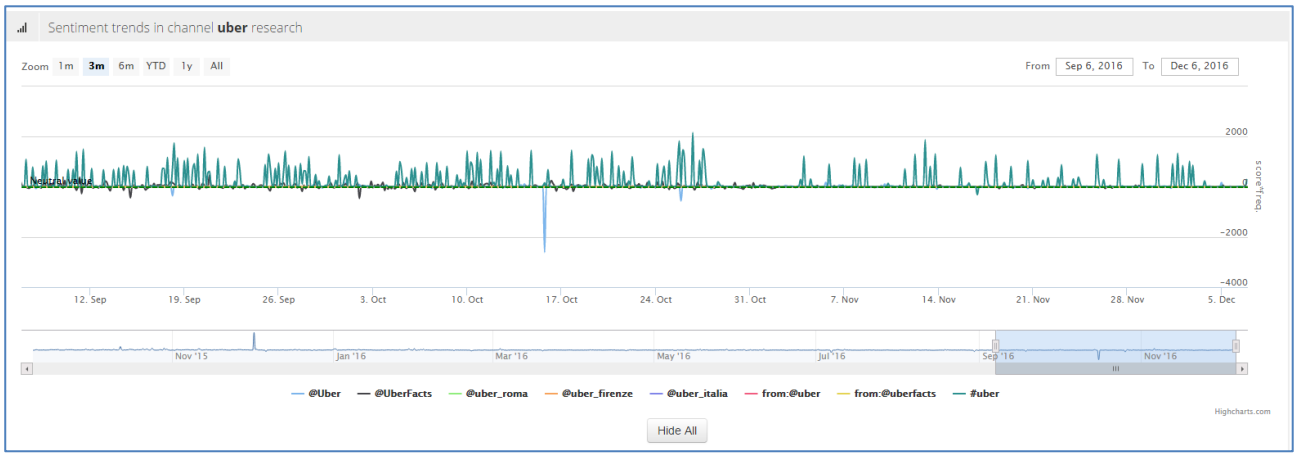


Figura 30: Sentiments trend in UBER researches

Figura 31: Sentiments resarch details

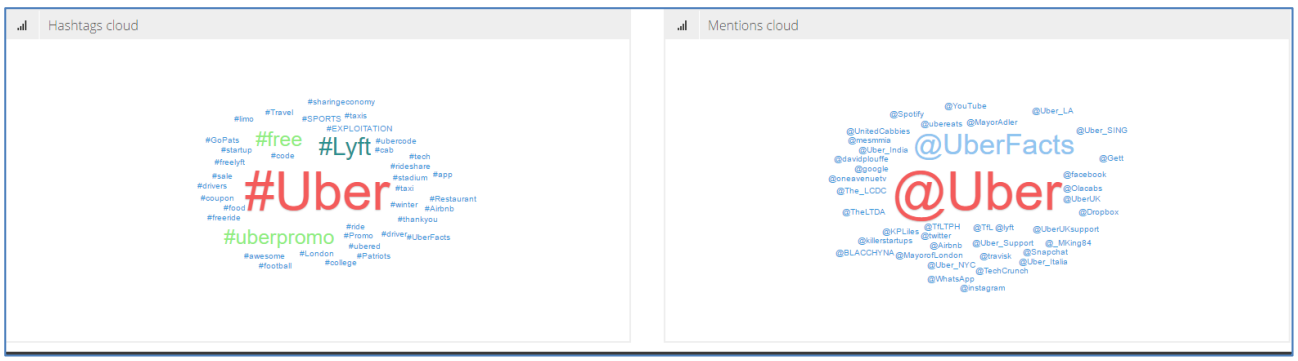


Figura 32: Word cloud

Figura 33: Sentiment analysis of single search

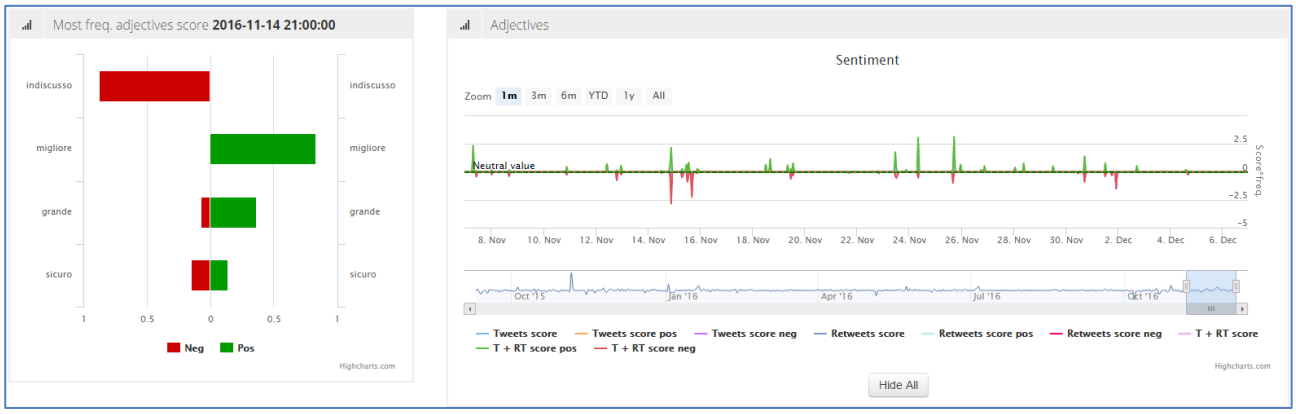


Figura 34 Sentiment per adjective with most frequencies per single search

Figura 35: Sentiment per verbs with most frequencies per single search

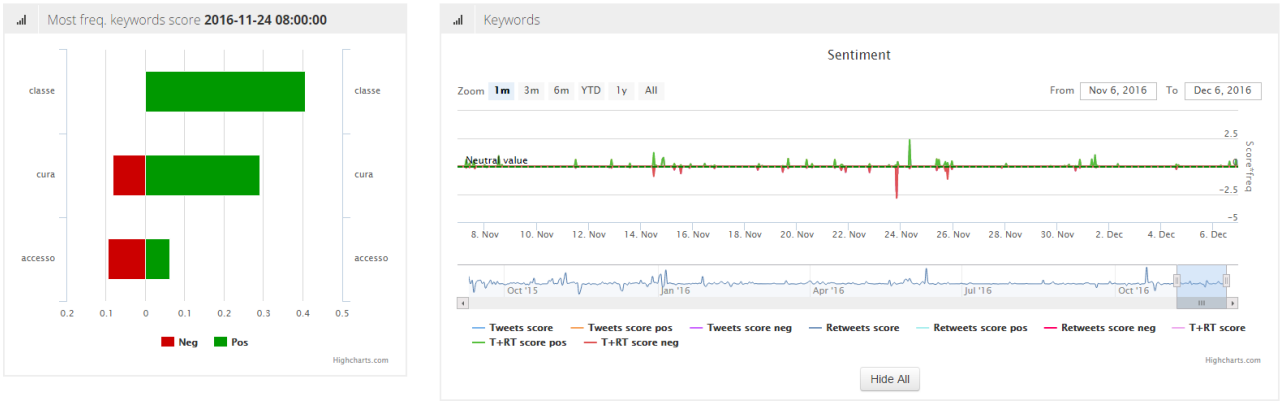


Figura 36: Sentiment per keywords with most frequencies per single search

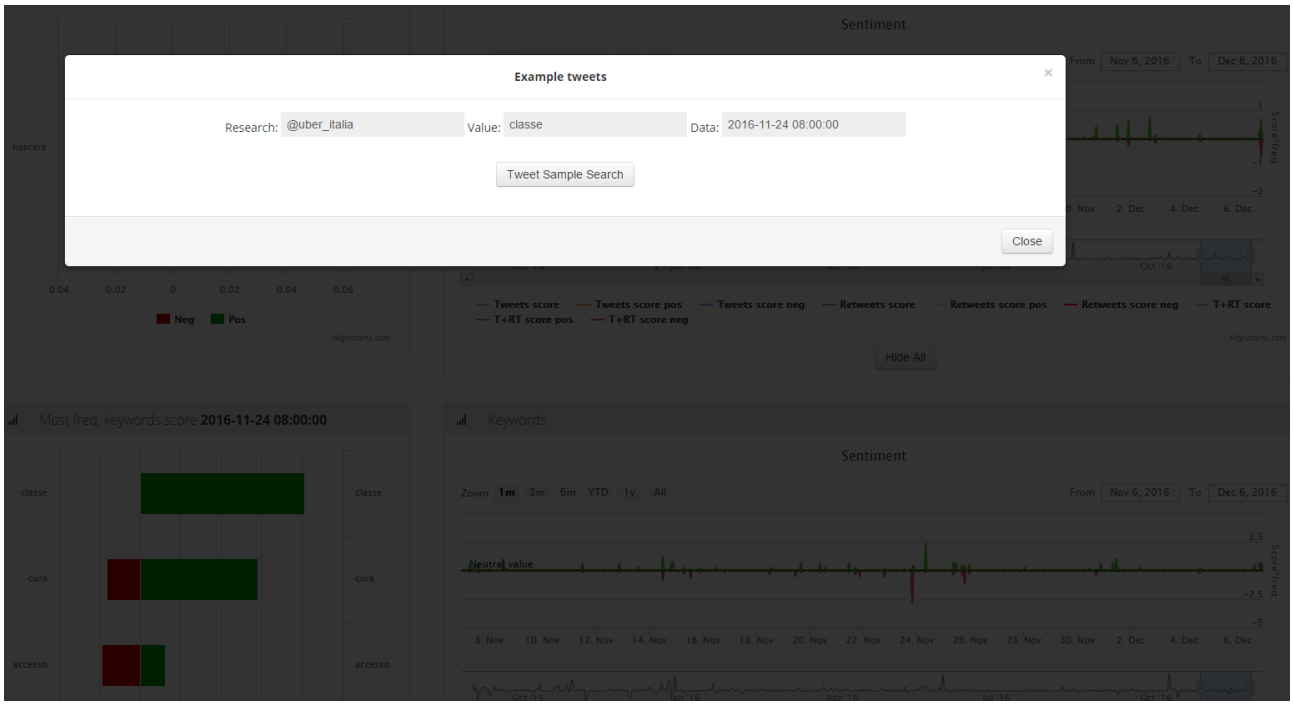


Figura 37: Advenced solr search

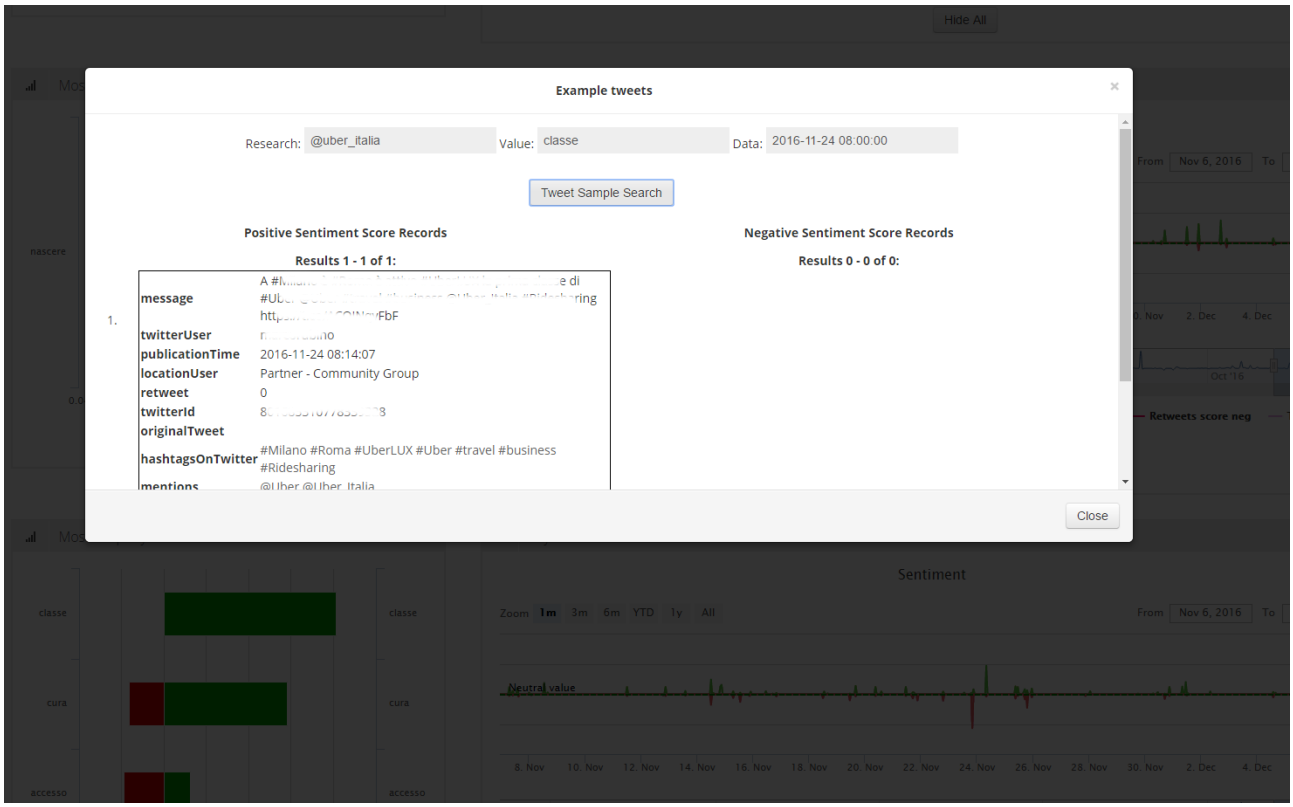


Figura 38: solr result

4 Bibliografia

[1] Ministero delle infrastrutture e dei trasporti: http://scioperi.mit.gov.it/mit2/public/scioperi/ricerca