DEMO PAPER: STAMAT: A FRAMEWORK FOR SOCIAL TOPICS AND MEDIA ANALYSIS

Giuseppe Serra, Thomas Alisi, Marco Bertini, Lamberto Ballan, Alberto Del Bimbo

Università di Firenze - MICC giuseppe.serra, marco.bertini, lamberto.ballan, alberto.delbimbo @unifi.it - thomas.alisi@gmail.com

ABSTRACT

Analysis of trending topics in social networks, based both on textual and multimedia content, can be used by content providers to measure the buzz around their channels, and even to create new material based on the current memes propagating across Twitter, Google+, Facebook etc.

STAMAT is a framework designed to provide a set of tools for social media analysis, and to create, through content curation, personalized web sites and magazines.

Index Terms— Social media analysis, topic detection, social networks.

1. INTRODUCTION

In recent years there has been an explosion of the usage of social networks that provide microblogging services that let users write short updates containing text and links to multimedia content, like Twitter and Google+. These tools allow people to easily produce content, thanks to the short length of texts that does not require an excessive effort in their creation.

Sharing of URLs/information and news reporting [1] are the main intentions among users. User interactions and participation to discussions regarding trending topics show that services like Twitter may be considered as a vehicle for news [2, 3]. Moreover the social aspect of conversations allows to consider human interactions as a channel that can be exploited to identify situations ranging from epidemics to flash mobs, joining social media with additional heterogeneous sources of data [4].

STAMAT (Social Topics and Media Analysis Tool) is a framework for the analysis of real-time social media, like Twitter, Facebook or Google+, taking advantage both of textual and multimedia content. The goal is to let users create an environment that automatically selects the most relevant news, links and images related to a set of topic or news of interest. STAMAT can be used for personal content curation, i.e. creating a personalized magazine from various news sources, or to let content producers and distributors to understand reactions of social networks to current news, such as feeds published by New York Times or TechCrunch. Laurent Walter Goix, Carlo Alberto Licciardi

Telecom Italia - Innovation and Industry Relations - Research & Prototyping laurentwalter.goix, carloalberto.licciardi @telecomitalia.it

2. DEMO

The demo shows the functionality of the system, implemented as a web application analyzing a set of data sources and displaying the outcomes of data processing. Users can select different news channels, assigning them to semantic categories; a pre-processing step starting from RSS feeds leads to web scraping, then HTML pages are analyzed to extract topics, named entities (see Fig. 1) and multimedia content. This information is used to retrieve related content on social media, providing both complementary information for news and a measure of the popularity and influence of news content (see Fig. 2).

A CBIR approach is used to evaluate how media embedded in news sites is propagated across social media; this can be used for different tasks, like selecting images that can be considered as representative of popular topics or evaluating if some media content is becoming viral [5].

3. THE SYSTEM

The system has been developed as a web application: the backend and analysis services have been developed in Java using the Play framework, while user facing web applications have been implemented with Backbone.js, Twitter Bootstrap and CodeIgniter, to provide a snappy user experience.

News items from RSS feeds and URLs are processed to determine their language using n-grams and Naïve Bayes classifiers, since this technique has proved to be effective also on short text fragments like tweets, to perform appropriate stemming and stop-word elimination; their content is then summarized using latent topics extracted with LDA [6,7] and named entities obtained using a mixed approach that employs a set of rules combined with gazetteers [8] and CRFs [9]. Topics and entities are used to perform an initial selection of relevant tweets that could be associated to news. Tweets are used to provide additional social information related to news elements, and also to re-rank news based on their influence within social network: the idea behind our ranking algorithm is that tweets can "vote" for news if they are similar to a news

ANAGE FEEDS			SELECT VOCABULARY	
ANAGE FEEDS	URL	add feed	SELECT VOCABULARY	
cache all feeds to database			categoria res	
his operation is pr	erformed periodically b	iy a cron job		finanza homepage politica teonologia
Show 30 c er	ntries		guotidieni	
Search:			corriere repubblica	sole 24 ore
id title	tags	ut	untrine	
	corriere, homepage	http://xml.comereobjects.t/nss/homepege	ami	
10 repubblica home	repubblica, homepage	http://www.repubblica.it/ss/homepage/rs	add selected tags	
11 repubblica cronaca		http://www.repubblica.it/rss/cronaca/rss2	0.xml COLOUR CODES	
12 repubblica	model politica.	http://www.repubblica.it/rss/politica/rss2.0		ard child
politica 13 sole, prima	repubblica, politica	http://www.isole24ore.com/res/primapag		
pegina	homepage			
		http://www.isole24ore.com/ss/finanza-e mercati.xml		
		http://www.androidworld.it/feed/		
16 Mashable Showing 1 to 8 o		http://feeds.mashable.com/mashable		
00				
politica delete selected tag	5			
iekoom halls Lab — 62	keywo Ap Androi	T Fopics and Media Analysis Tool rds • ple New York Photogra d (operating system) London Dai M	es arre vocabilities topic & entities topic apply Samsung Google Micros tan Los Angeles New York City Apple Icc.	towsty, schrinkersent + oft Music Kate Moss Sen Francisco Pome Loukiana Facebook Design
ekcon tali Lat 92	Stand Social Ap Androi Televia	Topics and Media Analysis Tool rds pie New York Photogra d (operating system) London Dail M icon Massachusetts Twitter Media Nation	aphy Samsung Google Micros ian Los Angeles New York City Apple Inc. I menuication (bee-able frod Pass Technology So Benuccompro Josef Songer Song Date Extrans. Inc.	howly, sameatestri + oft Music Kate Moss San Francisco Rome Louistana Facebook Dereign almeat metagony Viss Divid reactive offer
ekcon talis Lat 02	Stand Social Ap Androi Televia	Topics and Media Analysis Tool rds ple New York Photogra d (operating system) London Del M ion Massachusetts Twitter Media loc	aphy Samsung Google Micros ian Los Angeles New York City Apple Inc. I menuication (bee-able frod Pass Technology So Benuccompro Josef Songer Song Date Extrans. Inc.	howly, sameatestri + oft Music Kate Moss San Francisco Rome Louistana Facebook Dereign almeat metagony Viss Divid reactive offer
elecore table Late 202	Stand Social Ap Androi Televia	Torpics and Media Analysis Tool rds pie New York Photogra d (operating system) London Dal M ion Massachusets Twitte Media co Nationa as Boogra Eure Application othere Si	aphy Samsung Google Micros ian Los Angeles New York City Apple Inc. I menuication (bee-able frod Pass Technology So Benuccompro Josef Songer Song Date Extrans. Inc.	howly, sameatestri + oft Music Kate Moss San Francisco Rome Louistana Facebook Dereign almeat metagony Viss Divid reactive offer
iekon hala Lak 82	Straw Social Ap Androi Televis News m people	T Copies and Mode Analysis Tool rds • DIE New York Photogra d (operating system) London Dai M in Massechuster Statter Merice National St	tophy Samsung Google Micros Itan Los Angeles New York City Apple Inc. In mencentarity David Parts Tarritage for Banga cargony David Parts Tarritage for Banga cargony David Parts Tarritage for Banga Cargon Sams Cargony Cargony Cargony Cargony Cargonizations Cargony Cargony Cargony Cargony Cargonizations Cargony Cargony Cargony Cargony Cargonizations Cargony	hrve, ustrement * off Music Kate Mose Sen Francisco Tem Latera Restors: Dergen ansis Respon to Version Restor Serve Server Sectors (spectra Decetors e
vienem hale Lat 0	Straw Social Ap Androi Televis News m people	T Capace and Media Analysis Tool rds ● D O New York Photograp of (perening system) London Del M In Massachusetts Turter Meano the Appendix on Appendix on the Meano New Appendix on Appendix on the Meano New Appendix on Appendix on the Meano Del Focus Nell	In Los Angeles New York City Apple for. That Los Angeles New York City Apple for. The City Apple for the Interference Strate and	the second of th
vecon tale Lai 0	Straw Social Ap Androi Televis News m people	Topos and Moda Analysis Tool rds ple New York Photogra generating years topost and the generating years topost anges for Apparetanements machine mach	Apply Samsung Google Micros in Los Argenes New York City Apple Inc. I manuscript Search and New Insteading the Search and the Search and Search the Instead Search and Search and Search Comparison and Search and Search Amazon Media Lab Limited	torig advances * off Music Kate Moss San Francisco San Contacto Marcine Contacto San Contacto Marcine Contacto San Contact San Co
Nean tale La 0	STAMM Social Ap Andrai Televia News in people Do	Topics and Madia Analysis Tool "Gips and Madia Analysis of employee New York Photogra- di (persting system) London Del M for Massachusetis Tahler Wals Magnetic Analysis Kananya Kanada Wayne Kananya K	In Los Angeles New York City Apple for. That Los Angeles New York City Apple for. The City Apple for the Interference Strate and	the second of th
Necentralis La 0	STAMM Social Ap Andrai Televia News in people Do	Topora and Mada Analysis Tool Space and Mada Analysis Tool Space A	han bar and the second	And advances - Contract of Music Kata Mose Ser Francisco Music Anter Antonico Music Antonico Ant
Glean tale Lat 9	STAMM Social Ap Androi Televia News In people Do	Topics and Media Analysis Tool Copies and Media Analysis (coprating system) (coprating system) (copies to so and there is a solution (copies to solution and the media (copies to solutio	Apply Samsung Google Micros tan La Agers fee Yor Chy Agets Inc. mensated fees also read fees that the development of the transmission development of the transmission of the transmission of Amazon Media Lab Limited New York Times Anna Cartonte Campras Mobile Tech News	Andreas and
Weam hale Late 10	STAMM Social Ap Andrai Televis News In People DO Thoma	Toposa and Mada Analysis Tool " " Toposa and Mada Analysis Tool " " Toposa and Mada Analysis Tool " " Toposa Analysis Tool " " " Toposa Analysis Tool " " " " " " " " " " " " " " " " " " "	aphy Samsung Google Micros tan LorAques Ner York Ciy Agel Inc. Microsoft Circle Circle Circle Circle Microsoft Circle Circle Circle Circle Annacon Media Lab Limited New York Times Anna Cardens Campana Mobile Tech News Mobile Tech News Mark Art Print Finns	Andreas and
Alcentals Lat 9	STAM Social keywo Ap Andron Televis News re people Do Thoma Ha Covers Ha	Topics and Media Analysis Tool Copies and Media Analysis (coprating system) (coprating system) (copies to so and there is a solution (copies to solution and the media (copies to solutio	phy Samsung Google Micros an Langues Mar Har Of A desired in the second second second second second manual second second second second and second second second second second and second second second second second Marcha Lab Lanted New York Times Marcha Lab Lanted New York Times Machine Tech Newsen and Andrea Lab Lanted New York Times Machine Tech Newsen and Andrea Lab Lab Laborations And Andrea Lab Laborations Andrea Lab Laborations Andrea Laboratio Andrea Laboratio Andrea Laborations Andrea Laboratio	Andreas and
elecentrali Lai 🕈	STAMM Keywo App Androi Televe People Do Thoma Thoma Cove to Zaza Pir Present	Name of the second seco	phy Samsung Google Micros the Longenetic Net York Dy Agelitic. International Internation	Andreas and
fakar lak di	STAMM Social Standard Regions App Andreas Televis In Televis Intel Inte	To a series and Media Analysis Tool rds • ple New York Photogra dispersing system London Des no Messelutes The The Series on Messelutes The The Series on Messelutes The The Series on Messelutes The Series Des New York Photographics and Control Series and Control Messelutes The Series of the Series Series Series (Control Messelute Series) (Control Mess	phy Samsung Google Micros to Langevin Ner Yin Gy Agel Io. I manual and the second second second manual second second second second and second second second second and second second second second Madia Lab Limited New York Times Anna Gardnesh Campras Mobile Tech News Mobile Tech News Mobile Tech News Mobile Tech News Pholography Design 1 iferty Automotive Food Home	Andreas and
	STAMM Social Standard Regions App App Andres Television Standard S	The search of Mode Androps is Tool **********************************	aphy Samsung Google Micros man Lon Agenes Ner Yok Dy Ageletic. International Annual Part Internet Sectorsmann, Gener Mark, Landon Carl Annual Carl Market Carl Comparison Conference Commands Model La Lumindo New Yok Times Model La Lumindo New Yok Times Tech News Arna Carl Annual Carl Sector Model La Lumindo Mobile Tech News Photography Design I Infertyle Automotive Food Home	Andreas and
	STAMM Social keywo Ap Ardrou Telowiew Do Telowiew Zas IV Teores Come In Come I	The second secon	phy Samsung Google Micros to Langevin Ner Yin Gy Agel Io. I manual and the second second second manual second second second second and second second second second and second second second second Madia Lab Limited New York Times Anna Gardnesh Campras Mobile Tech News Mobile Tech News Mobile Tech News Mobile Tech News Pholography Design 1 iferty Automotive Food Home	Andreas and

Fig. 1. STAMAT: *top*) managing news feeds; *bottom*) managing topics, entities and concepts.

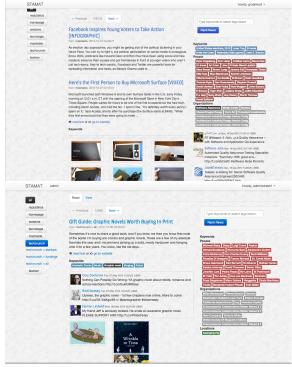


Fig. 2. STAMAT: *top*) news and social media associated to topics and entities extracted from all the news; *bottom*) news and social media associated with a single news item.

title, by computing its similarity with TF-IDF, after stop-word elimination.

Media elements embedded in news and tweets are indexed using a set of global and local features (CEDD, MPEG-7 descriptors, SIFT BoWs). These features are used to perform CBIR using news and social multimedia; news images are used to search for similar images that propagate across social channels (e.g. to evaluate if some fashion photos hac an impact on a social network, see Fig. 3), while social media are used, similarly to tweets, to "vote" the most representative images of a news topics. To speed up search, images are indexed using an approximate search data structure based on inverted files proposed in [10].

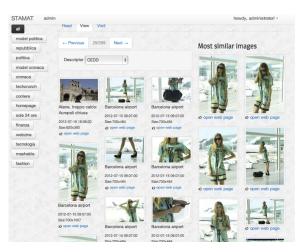


Fig. 3. Example of search of fashion images in websites and social media.

4. REFERENCES

- A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc.* of WebKDD and SNA-KDD, 2007.
- [2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proc. of WWW*, 2010.
- [3] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks," in *Proc. of ICWSM*, 2010.
- [4] V. K. Singh, M. Gao, and R. Jain, "Situation recognition: an evolving problem for heterogeneous dynamic big multimedia data," in *Proc. of ACM MM*, 2012.
- [5] L. Xie and H. Sundaram, "Media Lifecycle and Content Analysis in Social Media Communities," in *Proc. of ICME*, 2012.
- [6] D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012.
- [7] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu, 2002.
- [8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proc. of ACL*, 2002.
- [9] J.R. Finkel, T. Grenager, and C. Manning, "Incorporating nonlocal information into information extraction systems by Gibbs sampling," in *Proc. of ACL*, 2005.
- [10] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in *Proc. of InfoScale*, 2008.