# Towards a Big Data Solution for Analyzing the Reputational Impact of the Tennis Italian Open

Ermelinda Oro, Massimo Ruffolo
National Research Council (CNR)
Altilia srl
Via P. Bucci 41/C, 87036
Rende (CS), Italy
{oro,ruffolo}@icar.cnr.it
{linda.oro, massimo.ruffolo}@altiliagroup.com

## ABSTRACT

The Italian Open is part of the nine main tennis tournaments in the world after those of the Grand Slam. This event has grown over the years generating value, not only for the event itself, but also for social and economic context in which it is grafted, and in particular to the city of Rome that annually hosts it. In this paper we present the initial results of an analysis that aims at evaluating the impact of the Italian Open by measuring its reputation on the web. To get our goal we exploit a Big Data Analytics platform that enables to: (i) extract contents from heterogeneous sources, (ii) perform text analysis, and (iii) visualize results. In particular, we carry out a quantitative and qualitative analysis of texts related to the event coming from social networks and newspapers. The work is a step forward in the analysis of these kind of tennis events not considered until now, and an advancement towards the definition of a shared method to process online resources about sports events.

## Keywords

Text analysis, Social media analysis, Sentiment Analysis, Smart Data Analysis, Tennis Italian Open, Big Data Platform

## 1. INTRODUCTION

The Italian Open is part of the nine major tennis tournaments in the world after those of the Grand Slam. This event has grown over the years. Most important worldwide tennis players take part to this tournament that attracts a lot of attention generating value for the city of Rome, which get many social and economic benefits from the event.

In this paper we present preliminary results obtained by applying our methodology and initial analysis that aim at evaluating the impact of the Italian Open by measuring the reputation of this event within Italian social and online media. To get our goal we perform a quantitative and qualitative analysis of texts related to the event coming from social networks and newspapers by using a Big Data Analytics platform that enables to extract contents from heterogeneous sources, to perform text analytics, and to visualize results.



*KDD '16 August 14, 2016, San Francisco, CA, USA*
*Workshop on Large-Scale Sports Analytics*



The analysis of reputational impact will aim at identifying not only the volume and type of generated conversations and articles, but it will also aim at offering a more focused picture of which are the factors that impact on the reputation. Such a result will enable to identify suggestions capable to increase economic and social benefits.

In this work, the analysis of reputational impact aims to answer the following questions:

- When and how much do newspapers and social networks talk about Italian Open?
- What is the most discussed topic?
- What is the opinion about Italian Open?
- What are the most active users?

This article is organized as follows. Section 2 briefly describes the related work. Section 3 introduces the methodology and tools used for the analysis. Section 4 presents our initial results. Section 5 concludes the work.

## 2. RELATED WORK

Work related to the method we use to evaluate the impact of Italian Open falls in three different areas: (i) Text analytics and in particular sentiment analysis methods. (ii) Reputation analysis of sport events. (iii) Big Data Analytics platfroms.

Sentiment analysis returns the overall opinion of a text or document for one issue. Opinions are classified as positive, negative, or neutral. This kind of labeling can be used to summarize the content of opinionated texts and documents. A wide variety of features can be necessary for opinion and polarity recognition [13]. Sentiment analysis on social networks messages is attracting a lot of interest [2]. Pang et al. [14] provided a broad overview of some machine learning techniques used in sentiment classification. In [9] sentiment classification techniques are divided into machine learning, lexicon based and hybrid approaches. Walaa et al. [10] provides a recent comprehensive overview of the sentiment analysisin text mining field. The goal of this paper is not compare and choose the better sentiment analysis method, but just experiment which kind of interesting information can be carried out about the Italian Open event.

Different studies have been conducted to explore the impact of hosting large-scale sport tourism events focusing on mainly tangible out-comes (i.e., economic benefits) rather reputational impacts [6]. Many works examined residents' perceptions, for FIFA World Cup [8], for resident of Naples [3]. In [16] the Facebook pages were examined, without using an automatic method, as a brand-management tool in college athletics. In [17] tweets from U.S. soccer fans during five 2014 FIFA World Cup games are collected.

They used sentiment analysis to examine U.S. soccer fans' emotional responses in their tweets, particularly, the emotional changes after goals. Alves et al. [5] compared two approaches for sentiment analysis of Portuguese tweets related to the FIFA's Confederations Cup. In [1] the effectiveness of a machine learning method for providing positive or negative sentiment on tweets around sporting events are examined. To the best of our knowledge, there not exist papers analyzing the impact of the Italian Open event.

Big Data arises with many challenges, such as difficulties in data capture (in particular, from unstructured sources), data analysis and data visualization. In [4] a state-of-the-art of techniques and technologies recently adopted to deal with the Big Data problems is presented. A technology independent reference architecture for big data systems is presented in [12]. A complete big data platform needs different type of technologies for enabling data acquisition, processing, analysis and visualization. The paper [7] review relationship between big data and cloud computing. Cloud computing is a powerful technology to perform massive-scale and complex computing. In cloud based systems, as discussed in [15], orchestration is fundamentally important to improve scalability, and in particular, workflow orchestration is needed for automatically combining and connecting different data sources. In [11] we presented the MANTRA Smart Data Platform that makes use of contextual workflows and text analysis method for semantically extract data from documents. We chose to use our platform because it enables us to implement the whole presented method, from the extraction to the analysis, by using a single interface. In addition, MANTRA enables us to test different methods that can be put in the platform in a very simple wa by adding new APPs.

## 3. PROPOSED METHOD

In the last decade, the Italian Open (also called IBI) recorded an extraordinary growth trend, with a higher increasing of sold tickets. But such a measure is not the only one that should be considered. In fact, beside the significant growth in size (the public), IBI has grown also in media visibility (including social networks and online newspaper). In last years IBI has been the main individual event for attractiveness in Rome, which annually hosts the event. Therefore, it is important to understand the reputational effects: knowledge about the high number of conversations that are generated online on IBI, and the quality in terms of content, emotions, and the different results based on the different involved media.

This paper describes the results of the analysis of contents about the IBI event published, during 2014 and 2015, in articles of tennis newspapers, and in posts (and comments) of the official Facebook page of IBI. One of the objectives is to extend this work considering the 2016 data in order to observe annual trends. To perform the analysis, we used the heterogeneous sources listed in Table 1 below. The dataset, created by automatically extracting contents from these sources, is publicly available[1].

To implement the proposed solution, we used our MANTRA Smart Data Platform [11], that combines semantic, big data, and cloud computing technologies. Therefore, it provides, in an integrated way, functionalities to create Big Data Analytics applications by exploiting a method based on contextual worksflows and APPs. The user can: (i) capture both structured and unstructured data from enterprise systems and the web, (ii) augment data by using NLP and semantic features, (iii) harmonize data with respect to the analytical purposes, (iv) perform analysis, (v) visualize data and analysis results by using reports and charts that help sense making.

The analysis of IBI related contents consists of two main steps,

[1]Datasets are available at: http://www.lindaoro.com/datasets/ibi.html

**Table 1: List of considered online sources**

| Fonte | Sito Web |
|---|---|
| FB Internazionali | Internazionali-BNL-dItalia-267627913872 |
| Livetennis | www.internazionalibnlditalia.com |
| Ubitennis | www.livetennis.it |
| Tennisitaliano | www.ubitennis.com |
| Spaziotennis | www.tennisitaliano.it |
| Tennisbest | www.spaziotennis.com |
| Tennis | Tennisbest.com |
| Federtennis | Tennis.it |
| | www.federtennis.it |

described below, defined in visual way by using the MANTRA Workflow Modeler and the MANTRA Mashboard Modeler.

The MANTRA Workflow Modeler (see Fig. 1) allows user to design workflows where each block/step is a MANTRA APP, i.e. a software module that internally performs a complex computational task. MANTRA APPs can be distinguished according to their purpose that can be: acquisition (e.g., automatic content extraction from the web, connection to social network APIs); normalization and transformation (e.g., data fields cleaning and standardization); analysis (e.g. entities and relationships extraction, sentiment/opinion analysis).

The MANTRA Mashboard Modeler allows for creating dynamic dashboards in which the generated charts can be navigated and explored along different dimensions by combining and aggregating various types of collected data.

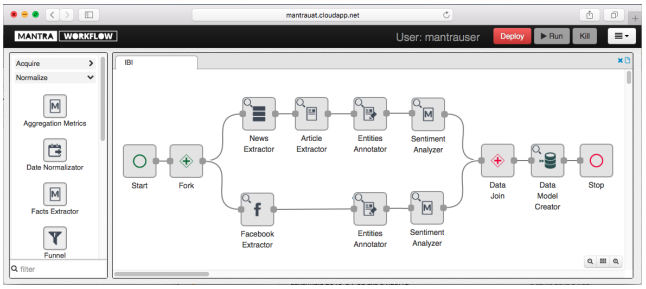


**Figure 1: An Example of MANTRA Contextual Workflow. It enables to: (i) acquire clean articles from online newspapers, posts and comments from Facebook; (ii) extract entities; (iii) analyze sentiments; (iv) create the enriched dataset that can be visualized and explored by the MANTRA Mashboard Modeler.**

The proposed methodology involves three main stages:

**Data Acquisition**. By considering the Fig. 1, the "News Extractor" step takes as input the home page URL of online newspapers and recognizes links of articles in the page. Then, the "Article Extractor" step extracts the clean content, title, images, and date of each linked article in a structured format. The "Article Extractor" method is implemented by using an euristic method, which exploits both visual/spatial and DOM features. We extracted articles and posts/comments about the IBI events held in 2014 and 2015. For each analyzed website are extracted only news that speak about IBI. Such a selection is performed giving terms and their synonyms that the article should contain as input of the APP. We intent to apply deeper semantic methods. In this initial study were used and combined simple terms like "IBI", "Master (s)? 1000 (a)? Rome", "Foro Italico", "International (BNL)?", "WTA (Primier|Master)? (5)? (of)? Rome". The "Facebook Extractor" APP allows for extracting posts, and comments in the official Facebook

page of the IBI by using the social network's API. Different features are extracted, e.g,: message text, shared links and images, number of received likes, number of shares of the post, relationship between messages and comments.

**Extraction of the Entities and Opinions**. By using the "Entities Annotator" and "Sentiment Analyzer" steps, shown in Fig. 1, we identify entities and expressed opinions in news and messages acquired from the web. In particular, we applied Natural Language Processing and sentiment analysis capabilities implemented in the corresponding MANTRA APPs. The taxonomy adopted to annotate entities of interest is represented by using the MANTRA Language, i.e. a first order logic based language that enables to represent domain knowledge. The type and quality of information that the machine automatically recognizes, are based on the modelled domain knowledge. In this initial work, we represented players, sporting events, sports organizations, hashtags and emoticons (relevant for Facebook). The sentiment analyzer is able to recognize, not just the positive or negative polarity, but also the intensity of the polarity and the opinion-target (i.e. the entity target of the opinion).

**Data Visualization and analysis.** This is achieved by creating Mashboards that enable detailed analysis, showing visual and concise results to business users. News and post/comments are analyzed considering both quantitative information (e.g.,: number of news, various types of sources, different entities, etc.), and qualitative information to understand the type of shared contents by users, mentioned entities, opinions that derive from posts, comments and articles. Some examples of the performed analysis on IBI are shown in the following section.

## 4. ANALYSIS RESULTS

In this subsection we show some results obtained by using the MANTRA Mashboard that enables queries and visual aggregations on data. The analysis of the reputational impact on articles and posts/comments aims at answering the following questions: i) When and how much do newspapers and social networks talk about Italian Open? ii) What is the most discussed topic? iii) What is the opinion about Italian Open? iv) What are the most active users?

### *When and how much do newspapers and social networks talk about Italian Open?*

Fig. 2 shows how the news that talk about the IBI are distributed over the time.

**Figure 2: Distribution of IBI's news over the time**

The number of articles about IBI has been substantially increased in 2015 compared to 2014. The news extraction process for the 2016 is partial, because the event is not yet held when the paper has been written, but we can predict the same incremental behaviour. Obviously, there is much talk of the IBI in May and in the months preceding and following the event. The histogram shown in Fig. 3 shows the websites that published the largest number of news. The figure shows that "ubitennis" contains a lot of news about the IBI. The number of news displayed in the time graph does not match the number of news shown in the histogram because some articles

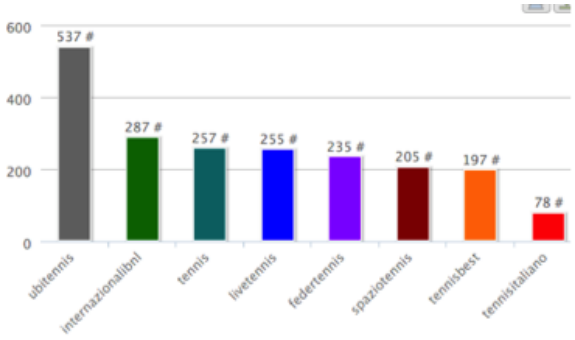


**Figure 3: Number of IBI's news per newspaper**

do not contain the pubblication date anymore. Likewise, analyzing Facebook fanpage we detected the increasing number of posts and comments, and also the numbers of users increased significantly over the time.

### *What is the most discussed topic?*

To analyze the content of articles and posts/comments, we use the taxonomy that we have defined by using the MANTRA Language. The news articles can be analyzed considering the title and the body of the article. As expected, the most frequent concept were the event of the IBI, the place where is played, and the players who disputed the matches, and in particular the most famous players in the international and national context. In Fig. 4 are shown the players that have been discussed more frequently in the analyzed years.


**Figure 4: Tag cloud about the "tennis_player" concept.**

In addition, the analysis has shown that newspapers discuss more about male events, whereas Facebook comments are often related to the female sex. Like in Twitter, hashtags and smiles are often used in the social network. Most commonly used hashtags in order of frequency are: #tennis, #ibi16, #wta, #atp, #ibi14, #ibi15, #countdown, #federer, #ajdenole, #masha

### *What is the opinion about Italian Open?*

In this subsection is shown the sentiment detected in the articles and in the post/comments. Our method is able to recognize, not just the positive or negative polarity, but also the intensity of the polarity and the opinion-target. Fig. 5 shows the percentage of articles that express positive, negative or neutral opinions. Normally, news articles express positive opinions. But, by analyzing individual sentences of news, negative statements can identified.

Same analysis have been performed on Facebook posts and com-

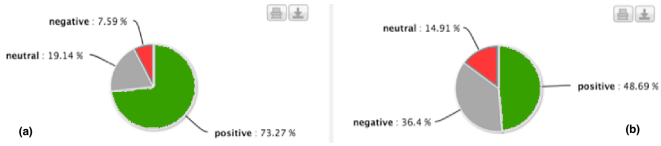


**Figure 5: Percentage of negative (red), positive (green) and neutral (grey) opinions expressed in the news considering (a) the whole article and (b) each single sentence.**

ments. We observed, that comments express more negative opinions than posts. This result is explained by the fact that posts are mainly informative messages. Clicking on negative sentiment, every sentiment and related opinion-target can be explored. For instance, by choosing negative sentiment, and selecting the opinion-target "Rome", we found that generally issues are about the inability to visit Rome.

### *What are the most active users?*

We extracted the most active users. We observed that the posts are essentially written by the International Tennis organizations, and only comments are written by Tennis fans. Future work will be focused on the social network analysis, such as influencers detection.

## 5. CONCLUSION

In this paper we described the initial results of the analysis aimed at evaluating the impact of the Italian Open by measuring the reputation of this event within social and newspaper.

Future work will be focused on implementing and applying machine learning techniques to identify new concepts in (semi)automatic way. We will extract further information, like "annotated facts" that characterize the event (i.e. triples that describe subjects, objects, and related actions). A comparison between natural-language-based and neural-network-based methods will be performed. Furthermore, we intend to analyze different input sources and contents (not only text, but also pictures and video), including 2016 data.

Finally, our final goal is a comparison with other events in order to identify the highest economic and reputational level that it can be reached. This comparison will allow the identification of factors that produce a better reputational impact. In particular, for the benchmark analysis we were identified three categories of comparable events: i) Events with maximum analogy (i.e., others events belonging to the ATP World Tour Masters 1000 circuit). ii) Events with medium analogy: these events related to different sports from tennis offer a useful comparison because of the importance and the impact they have on the territory of the host. iiiz) Events with low analogy (i.e. events not related to sports tournaments), but have an important impact on the hosting metropolitan areas. Such an analysis will provide precise directions and suggestions for actions and future investments.

## References

[1] P. Barnaghi, P. Ghaffari, and J. G. Breslin. Text analysis and sentiment polarity on fifa world cup 2014 tweets. In *ACM SIGKDD Workshop on Large-Scale Sport Analytics (LSSA)*, 2015.

[2] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.

[3] R. Caiazza and D. Audretsch. Can a sport mega-event support hosting city's economic, socio-cultural and political development? *Tourism Management Perspectives*, 14:1–2, 2015.

[4] C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[5] A. L. Firmino Alves, C. d. S. Baptista, A. A. Firmino, M. G. d. Oliveira, and A. C. d. Paiva. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, pages 123–130. ACM, 2014.

[6] H. J. Gibson. Sport tourism: a critical analysis of research. *Sport management review*, 1(1):45–76, 1998.

[7] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47:98–115, 2015.

[8] W. Kim and M. Walker. Measuring the social impacts associated with super bowl xliii: Preliminary development of a psychic income scale. *Sport Management Review*, 15(1):91–108, 2012.

[9] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *The semantic web: ESWC 2011 workshops*, pages 88–99. Springer, 2011.

[10] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[11] E. Oro and M. Ruffolo. Using apps and rules in contextual workflows to semantically extract data from documents. In *roceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, 2015.

[12] P. Pääkkönen and D. Pakkala. Reference architecture and classification of technologies, products and services for big data systems. *Big Data Research*, 2(4):166–186, 2015.

[13] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[14] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[15] C. Vitolo, Y. Elkhatib, D. Reusser, C. J. Macleod, and W. Buytaert. Web technologies for environmental big data. *Environmental Modelling & Software*, 63:185–198, 2015.

[16] L. Wallace, J. Wilson, and K. Miloch. Sporting facebook: A content analysis of ncaa organizational sport pages and big 12 conference athletic department pages. *International Journal of Sport Communication*, 4(4):422–444, 2011.

[17] Y. Yu and X. Wang. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans' tweets. *Computers in Human Behavior*, 48:392–400, 2015.