Interpreting Reputation Through Frequent Named Entities in Twitter

Nacéra Bennacer Seghouani, Francesca Bugiotti, Moditha Hewasinghage, Suela Isaj, Gianluca Quercini

Interpreting Reputation through Frequent Named Entities in Twitter

Nacéra Bennacer, Francesca Bugiotti, Moditha Hewasinghage, Suela Isaj and Gianluca Quercini

LRI, CentraleSupélec, Paris-Saclay University Gif-sur-Yvette, 91190, France
{firstname.lastname}@lri.fr

Abstract. Twitter is a social network that provides a powerful source of data. The analysis of those data offers many challenges among those stands out the opportunity to find the reputation of a product, of a person, or of any other entity of interest. Several tools for sentiment analysis have been built in order to calculate the general opinion of an entity using a static analysis of the sentiments expressed in tweets. However, entities are not static; they collaborate with other entities and get involved in events. A simple aggregation of sentiments is then not sufficient to represent this dynamism. In this paper, we present a new approach that identifies the reputation of an entity on the basis of the set of events it is involved into by providing a transparent and self explanatory way for *interpreting reputation*. In order to perform this analysis we define a new sampling method based on a tweet *weighting* to retrieve relevant information. In our experiments we show that the 90% of the reputation of the entity originates from the events it is involved into, especially in the case of entities that represent public figures.

Keywords: Reputation, Frequent Itemsets, Sampling, Opinion Mining

1 Introduction

Twitter is one of the most popular social media platforms. A user of Twitter can follow any other user in the social network and the higher number of followers he has, the more reachability his tweets have. Twitter also has mechanism to spread information by means of retweets and favorites. The more retweets and the more favorites a tweet gets, the more it spreads, as it gets more audience.

Given that any kind of information can be posted and shared, it is possible to filter out tweets related to people, products, organizations or any other entity of interest. Data crawling depends on the APIs provided by Twitter and retrieving relevant data is a challenge, due to the noise. The opinion about an entity held by the public is widely known as *reputation*. Natural language processing tools are quite efficient for extracting sentiments expressed in a single tweet and the overall reputation of the entity can be only calculated by a simple sum of the sentiments of the individual tweets it is involved into.

However entities are characterized by their dynamism: they collaborate and create events. Public figures offer the typical example of reputation influenced

by events. In this paper we show how the reputation of a person or a product can be construct as a combination of the events it is involved into. In our experiments show that 90% of the reputation of an entity comes from these events. We also propose a sampling method that is based on the retweets, the number of followers, and the favorites. We finally show that the weighted sample technique yields richer information and it is able to discover more events.

The remainder of the paper is organized as follows: in Section 2 we discuss the related work, we provide a detailed description of our approach in Section 3, we present our experimental results in Section 4 and we conclude in Section 5.

2 Related Work

Twitter has been used broadly for gathering information about an entity of interest. Finding the attitude (positive, negative, or neutral) towards a topic is known as sentiment analysis. Most of contribution in the field focuses in finding sentiments in the tweet-level [1, 4, 10], some of them suggest aggregating the sentiments as a simple sum [12, 15, 16], while the problem of the reputation of an entity has not been specifically addressed. Moreover, the usage of emoticons contributes to the sentiment [11]. In [10] they have used Internet specific acronyms, emoticons, and domain specific text processing to successfully detect the sentiments.

Machine learning techniques prove to be effective with sentiment analysis [15, 17]. In order to augment the accuracy of the classifier, Semantic Sentiment Analysis is used in [13]. Emoticons, repeated letters or acronyms have been used in [3]. Domain-dependent sentiment analysis has been studied in [18] and the effect of hashtags in assigning sentiment scores to tweets in [16]. Sentistrength has been developed using emoticons, repeated letters, phrasal verbs and everyday expressions, exclamation marks and repeated punctuation and it shows a higher accuracy compared to several other learning methods [14]. [5] uses pattern discovery and mining of comparative sentences inside blogs, forums, and product reviews. [12] uses a different approach where the entities are further classified into topics and the overall opinion is summarized on the different topics.

3 Approach

In order to extract reputation of people and products from Twitter, we describe the dataset in terms of the set of Named Entities that are defined in it. As first step we select an entity of interest and we collect data related to this entity by querying Twitter. As second step we enrich the information in the tweets using a sampling technique. On top of the sampled data, as third step, we apply frequent itemset mining technique to extract the Frequent Named Entities related to the entity of interest E. As final step, we extract the reputation of the entity E analyzing the collaboration of this entity with the other related entities.

Our goal is to interpret the reputation of an entity E, having as input the Twitter data, and as output defined in Equation 1 the reputation of an entity

E as a collection of reputations of the events it is involved:

$$R = \{(i_k, R_k) | 1 \leq k \leq n\} \quad (1)$$

where i_k is a frequent set of e_i and R_k is their reputation.

3.1 Weighted Sampling

The common problem that has been widely addressed is the data extraction. Twitter has provided a REST API[1] which allows running queries against the data to retrieve a sample of the actual content in Twitter. Retrieving tweets related to a specific entity is done by querying Twitter with a keyword. However this does not guarantee the relevance of the information. The most important question that we should pose is: "Do we went a statistically representative sample that aligns with the real large Twitter dataset or do we want a filtered sample that focuses on the relevant tweets?". Several papers have contributed to find a statistically representative sample [6, 8], while others highlight focused on crawling ([9]) or Expert Sampling ([7]). Inspired by the latest work, we consider three main parameters that influence the quality of the tweet: the number of times the tweet is retweeted, the favorite count of the tweet, and the number of followers of the user that has tweeted.

We sample the tweets according to these parameters. All the tweets are ranked taking into account: the number of retweets, the favorite count and the number of followers of the user. Then, a weight (w_i) is assigned to tweet t_i that comes from the average of all three rankings. We sample tweets according to their weight; we generate a random number (ϱ) of each t_i, if $w_i > \varrho$ than the tweet is selected to the weighted sample.

3.2 Reputation of Frequent Named Entities

The weighted sample S will be used in order to find the reputation of the entity, aided from the frequent entities in S. Named entities (NE) carry valuable information as they represent people, location, time, and monetary values. Considering t_i as a transaction containing a set of entities e_i as items, in addition to the traditional concepts of [2]: we consider a Frequent Named Entity (FNE) as a set of e_i that is maximal for a predefined support s in S. A FNE describes an event, which is usually associated by a reputation.

In this context the *reputation* of e_i, derived from S of k tweets will be the ratio between the sum of all positive sentiments and the negative sentiments. The sum of all underlying positive sentiments of tweets, as well as the sum of the negative ones can be transformed into normalized proportions that indicate the reputation of an entity ($\frac{\sum_{i=1}^{k} pos_k}{\sum_{i=1}^{k} pos_k + \sum_{i=1}^{k} neg_k}, \frac{\sum_{i=1}^{k} neg_k}{\sum_{i=1}^{k} pos_k + \sum_{i=1}^{k} neg_k}$). We propose finding FNEs and interpret the reputation of the entity of interest by the FNEs and their reputation. This approach is described in Algorithm 1. We intend to

[1] https://dev.twitter.com/rest/public

find the tweets that contain the FNEs and aggregate their sentiment (line 5, 6 and 7). In this way, the reputation of an entity can be explored analyzing its relations with other entities. This procedure is transparent to the user.

Algorithm 1 Calculating aggregated reputation of the FNEs

Input: A collection of tweets from S with their sentiment and set of entities $\{< t_i, E_i, pos_i, neg_i >\}$ where $E_i \leftarrow \bigcup_{j=1}^{n}, e_j \exists he_{ij}$,
A set of FNEs $< i_k >$ explored in S list
Output: Reputation R expressed as combination of $\{(i_k, R_k) | 1 \leq k \leq n\}$

1: **for each** FNE i_k **do**
2: $pos_k \leftarrow 0$
3: $neg_k \leftarrow 0$
4: **for each** $< t_i, E_i, pos_i, neg_i >$ **do**
5: **if** $i_k \subseteq E_i$ **then**
6: $pos_k \leftarrow pos_k + pos_i$
7: $neg_k \leftarrow neg_k + neg_i$
8: **end if**
9: $R_k \leftarrow (\frac{pos_k}{pos_k + neg_k}, \frac{neg_k}{pos_k + neg_k})$
10: add R_k to R
11: **end for**
12: **end for**
13: **return** R

4 Experiments

In this section we provide a description of the experiments we performed in order to evaluate our approach: we analyzed the richness of the samples, the effectiveness of frequent entity mining, and we compared the ranking of the sample to the population. To collect data about a certain topic we run a keyword query having as parameter a single string (such as '*Obama*'). We analyzed then the text of each retrieved tweet and we improved the quality of the text by separating merged words inside the hashtag. For example *#iamsohappy* and *#iam#sohappy* will be handled by our cleaning algorithm to produce "*I am so happy.*" We used the corpus of words of Sentistrength[2] for word identification and then different techniques for organizing the sentence and discarding not relevant words. We have used Stanford NLP[3] to identify the Named Entities from the retrieved tweets and Sentistrength as the sentiment analysis tool. Sentistrength scores a given text with a positive (from 1 to 5) and negative (from -5 to -1) values.

4.1 The datasets

We collected four datasets of tweets: *Obama* dataset, *Trump* dataset, *La La Land* dataset and, *The Voice* dataset by querying Twitter with respective strings. In the context of describing the characteristics of our datasets, we define two

[2] http://sentistrength.wlv.ac.uk
[3] https://nlp.stanford.edu

notions: *(i)* **Density of NE** - Expresses the average number of Named Entities linked to a tweet *(ii)* **Coverage of NE** - Represents the percentage of the tweets in the dataset that contain at least one Named Entity. The datasets regarding public figures have a high density of Named Entities, as well as a high coverage (Table 1). When it comes to the coverage and the density of Named Entities, *The Voice* is inferior to all other three datasets. For instance, compared to the public figures, it has half the density and half the coverage. It is interesting for our evaluation to take into consideration datasets with different characteristics.

	Obama	***Trump***	***La La Land***	***The Voice***
Density of NE	1.818	1.888	1.382	1.061
Coverage of NE	0.916	0.897	0.630	0.547

Table 1. Dataset characteristics related to NE

4.2 The richness of weighted sample

In our approach, we propose using weighted sampling for reputation discovery. We extract a random sample and a weighted sample from all datasets. We compare the richness of the information in terms of number of hashtags, number of URLs and number of NEs for 10 random samples and 10 weighted samples.

The average of the indicators are presented in Table 2. According to Table 2, the weighted sample is significantly richer in terms of the aforementioned indicators. Nevertheless, in terms of entities in *La La Land* and in terms of hashtags in *The Voice*, weighted sample has not been able to perform better. Since one of our parameters of interest is retweet count, sometimes for the movies and TV shows promotional tweets are retrieved, which might not be richer in information.

	Random			Weighted		
	Hashtags	Entities	URLs	Hashtags	Entities	URLs
Obama	14048.6	1828.5	5007.9	14256.1	1839.8	5230.2
Trump	8450.38	1609	2981.75	8655.25	1666.12	3094.5
La La Land	7986.9	1198.9	3102.9	9799.2	1081.6	3230.1
The Voice	1047.2	2856.7	1353	668.7	3368.2	1658

Table 2. Average indicators of the samples

4.3 Frequent Named Entity Mining in weighted sample

Frequent Named Entities are discovered through itemset mining techniques [2]. The tweets are considered as transactions and the NEs as itemsets. We used R to perform these experiments, arules package and eclat algorithm.

We used 50 random samples and 50 weighted samples to get an average of number of FNEs for each support value. For *Trump* dataset in Fig. 1 the weighted sample performs better for each of the support values, providing more FNEs than the random sample. *Obama* dataset in Fig. 1 shows a similar behavior as *Trump* dataset. For the same support, the weighted sample performs better, sometimes significantly better; in the low support values the weighted sample provides 20-40 more FNEs than the random sample. The weighted sample in *La La Land* (Fig. 2), in general, extracts more FNEs than the random sample. In the case of *The Voice* dataset (Fig. 2) the weighted sample is always superior to the random sample.

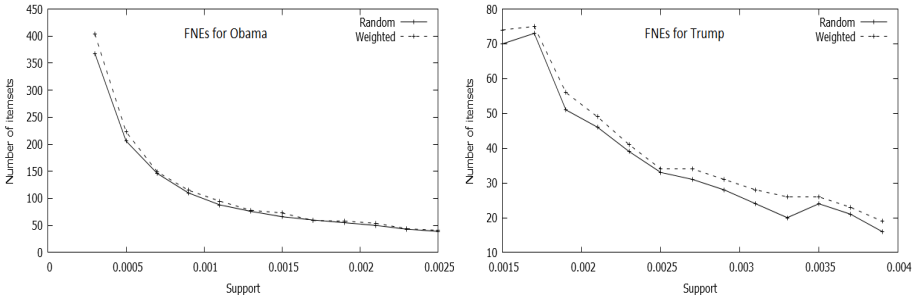

Fig. 1. Average number of itemsets for *Trump* and *Obama* datasets

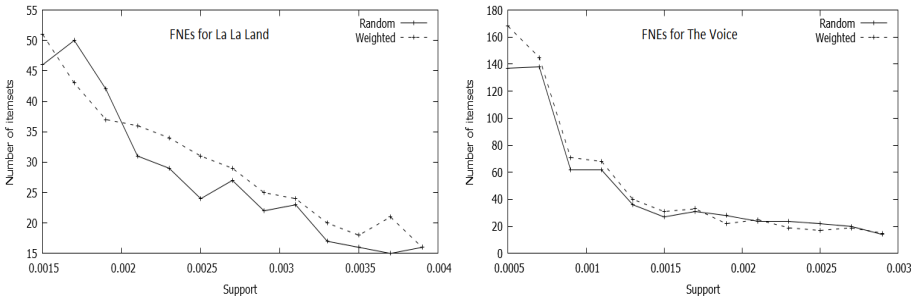

Fig. 2. Average number of itemsets for *La La Land* and *The Voice* datasets

4.4 Comparing the ranking of the FNEs

Since we are exploring FNEs through samples, we want to guarantee that the discovered FNEs are similar to the FNEs of the population. We ran the eclat algorithm on the whole datasets to discover the FNEs. We matched and ranked the FNEs in the population and in the sample. Then we calculated the Kendall

coefficient and the Spearman rank order for both rankings. We repeated the experiment for 10 samples from *Obama*, *Trump*, *La La Land* and *The Voice* dataset. The average values of 10 samples of each dataset showed a considerable similarity between the sample and the whole population in terms of ranking of itemsets; for *Obama*(0.79 and 0.89), *Trump*(0.76 and 0.60), *La La Land*(0.65 and 0.78) and *The Voice*(0.79 and 0.61) for Spearman and Kendall respectively.

4.5 Reputation through Frequent Named Entities

In this experiment we used Algorithm 1 and for each FNE, we found the sentiment and calculated its reputation. In order to respect the frequency of the FNE in the sample, we weighted the reputation by the support of the FNE. In the end, we calculated an overall reputation as $\sum_{k=1}^{n} R_k * s_k$,where R_k is the reputation of the itemset i_k and s_k is the support of the i_k in S.

We implemented this approach for all dataset, repeating the experiment 10 times for the weighted sample. The average accuracy is 90%. Both datasets related to public figures showed a precise alignment of the reputation explored through FNEs after weighted sampling with the reputation of the whole population. Nevertheless, in the case of the movie *La La Land*, we can distinguish a difference between both results, which comes from the fact that movies are not as dynamic as public figures, therefore, the reputation of a movie is enriched by FNEs, but not defined by them. In the case of *La La Land*, through Frequent Named Entities it is possible to discover viral events; in all of our 10 samples, the first FNE was related Emma Stone and JAEBUM and had a reputation of (+100, -0). It is important to note that our contribution does not focus on finding a reputation, but in enriching the interpretation of reputation by the means of Frequent Named Entities. This self-explanatory approach gives the user the possibility to interpret the information and since it breaks down the reputation of an entity into the reputation of the groups of entities it belongs to, the user has the freedom to use the pieces of reputation in a meaningful way.

	Obama	***Trump***	***La La Land***	***The Voice***
Whole population	(40.79,-59.20)	(32.04,-67.96)	(74.42,-25.57)	(56.06,-43.93)
Weighted Sample	(40.91,-59.08)	(38.22,-61.77)	(90.87,-09.12)	(55.28,-44.71)

Table 3. Reputation extraction through FNEs

5 Conclusions

We addressed the problem of reputation discovery and aggregation of sentiments by exploring the underlying entities that co-exist in the data. We introduced a weighted sampling technique to improve the richness of the dataset. We tested the power of Frequent Named Entity Mining on reputation discovery and we

showed that FNEs contribute to around the 90% of the reputation of the entity, especially in cases of public figures, who are highly dynamic in their collaborations with other entities. In this paper we used a ranking algorithm based on properties of interest to weight the tweets. Further studies on weighting techniques or choosing and transforming the properties of interest, could improve the quality of the sample.

References

1. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media, Association for Computational Linguistics*, pages 30–38, 2011.
2. R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216, 1993.
3. A. Bizhanova and O. Uchida. Product reputation trend extraction from twitter. *Social Networking, Scientific Research Publishing*, 2014, 2014.
4. J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM Conference*, 11:450–453, 2011.
5. X. Ding, B. Liu, and L. Zhang. Entity discovery and assignment for opinion mining applications. In *KDD Conf.*, pages 1125–1134, 2009.
6. M. Gabielkov, A. Rao, and A. Legout. Sampling online social networks: an experimental study of twitter. In *ACM SIGCOMM Conference*, pages 127–128, 2014.
7. S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi. On sampling the wisdom of crowds: random vs. expert sampling of the twitter stream. In *CKIM Conference*, pages 1739–1744, 2013.
8. M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: a case study of unbiased sampling of osns. In *Infocom*, pages 1–9, 2010.
9. G. Gouriten, S. Maniu, and P. Senellart. Scalable, generic, and adaptive systems for focused crawling. In *HT ACM Conference*, pages 35–45, 2014.
10. V. Hangya, G. Berend, and R. Farkas. Szte-nlp: Sentiment detection on twitter messages. In *SEM Conference*, volume 2, pages 549–553, 2013.
11. A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting emoticons in sentiment analysis. In *ACM Symposium on Applied Computing*, pages 703–710, 2013.
12. X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *KDD Conf.*, pages 379–387, 2012.
13. H. Saif, Y. He, and H. Alani. Semantic sentiment analysis of twitter. In *International Semantic Web Conference, ACM*, pages 508–524, 2012.
14. M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *ISI Journal, Wiley*, 62(2):406–418, 2011.
15. S. Van Canneyt, N. Claeys, and B. Dhoedt. Topic-dependent sentiment classification on twitter. In *ECIR Conference, Springer*, pages 441–446, 2015.
16. X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *CKIM Conference*, pages 1031–1040, 2011.
17. B. Xiang, L. Zhou, and T. Reuters. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *ACL Conference*, pages 434–439, 2014.
18. Z. Zhou, X. Zhang, and M. Sanderson. Sentiment analysis on twitter through topic-based lexicon expansion. In *ADC Conference*, pages 98–109, 2014.