Sentiment Analysis on Microblogs for Natural Disasters Management: A Study on the 2014 Genoa Floodings

Davide Buscaldi
Université Paris 13,
Sorbonne Paris Cité, LIPN, CNRS (UMR7030)
F-93430 Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr

Irazú Hernández-Farias
Pattern Recognition and Human Language Technology,
Universitat Politécnica de Valencia
Valencia, Spain
dhernandez1@dsic.upv.es

ABSTRACT

People use social networks for different communication purposes, for example to share their opinion on ongoing events. One way to exploit this common knowledge is by using Sentiment Analysis and Natural Language Processing in order to extract useful information. In this paper we present a SA approach applied to a set of tweets related to a recent natural disaster in Italy; our goal is to identify tweets that may provide useful information from a disaster management perspective.

1. INTRODUCTION

The number of users and messages in microblogs has been continuously growing in recent years, fostered by the spread of always-connected mobile devices. Microposts, the short messages published in microblogs, often report the status of users in a social or physical context, transforming the users themselves in real-time sensors about their local environment. Therefore, microblogging services like Twitter[1] have been proven to be an inestimable source of information for different tasks, such as opinion mining for commercial purposes [8], detecting social events like festivals or political unrest [6], and detecting natural disasters in real-time, for instance earthquakes [12, 10]. Event detection is usually performed by discovering unusual activity patterns, focused on a particular geographic area or on a given topic (usually specified by means of keywords). In order to carry out the detection tasks effectively, it is important to mine the right kind of information from the huge flow of posts (277,000 tweets per minute[2] all over the world). Detection results may vary greatly depending on the search terms, according to [12]. The purpose of a tweet may also vary depending on the user: [5] identified 4 different user intention types, varying from reporting news to conversation. Therefore, analysing the content of microposts may help to select only those that are relevant to a determined task. In the case of natural disasters, and more specifically in the case of disaster management, finding posts that indicate a situation of danger, worrying or generic alarm, may prove critical. In such cases, posts that report ongoing news, opinions, or even ironic comments are not particularly relevant and indeed may complicate the task of analysing the flux of information in such a critical situation.

Recently, there is a growing interest of the Natural Language Processing (NLP) research community on Sentiment Analysis (SA) or opinion mining, as testified by the new challenges in SA at different NLP conferences, such as SemEval [11, 7] or Evalita [2]. As a results of this research, some NLP methods and tools were adapted to work on microposts, overcoming a long-lasting gap due to the particular nature of the language used in these short messages. The results obtained in these challenges show that it is now possible to detect some types of irony and classify posts according to their polarity or subjectivity. In this paper, we formulate the hypothesis that we can use subjectivity, polarity and irony detection tools to filter effectively microposts related to natural disasters and therefore enhance the accuracy of the information available for the disaster management tasks. We suppose that subjective tweets are more important that objective ones because they are more likely to come from a person involved in the event rather than an objective tweet which is just reporting some news. We also suppose that ironic tweets are more likely to appear afterwards, for instance to criticize the response or blame the government. Finally, we suppose that tweets with a negative polarity are more probable to contain information about dangers or emergency situations in the context of a natural disaster. The rest of the paper is structured as follows: in Section 2 we describe the classification system used and the chosen scenario; in Section 3 we show the preliminary results obtained on the analysed data; finally, in Section 4, we draw some conclusions and discuss future works.

2. METHODOLOGY

Current SA systems used in polarity classification and irony detection are mostly based on machine learning approaches that exploit both surface features such as emoti-

[1] `https://twitter.com`
[2] `http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/`

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741727.

cons, exclamation marks and uppercase ratio, and lexicon-based features. Lexicons can be considered as affect dictionaries that map a word into a polarity score (positive, negative). For instance, SentiWordNet[3] [1] maps word senses (WordNet *synsets*) into a polarity score: "good" (first sense) has a positive score of 0.75, and "worst" has a negative score of 0.75. We participated in the Evalita2014 Italian SENTIPOLC (SENTIment and POLarity Classification) task with such a system, named IRADABE [3], obtaining one of the best results in the subjectivity (3rd with 0.6706 F-measure), polarity classification (2nd with 0.6347) and irony detection (2nd with 0.5415) tasks [2]. IRADABE relies on a Support Vector Machine (SVM) using surface and lexicon-based features. The lexicons have been adapted from English to Italian using machine translation. For the experiments presented in this paper, IRADABE was trained on the complete SENTIPOLC training+test set, consisting in 6448 tweets in Italian on various random topics, from politics to football. The dataset was POS-tagged using TreeTagger[4]. Here we describe the features used by IRADABE:

- *Bag-of-Words.* The most frequent words from the training corpus;
- *Emoticons frequency.* The frequency of emoticons expressing subjectivity, positiveness or negativeness;
- *Negative Words frequency.* Frequency of words that triggers negation (*mai* (never), *non/no* (not/no)), avversative conjunction or adverbs (*invece* (instead), *ma* (but));
- *URL information frequency.* The number of hyperlinks in a tweet;
- *Subjectivity features.* We took into account the presence of verbs conjugated at the first and second persons (those endings in "-o", "-i", "-amo", "-ate/ete") and personal pronouns ("io", "tu", "noi", "voi", and their direct and indirect object versions);
- *Tweet Length and Uppercase ratio.* The length in words of each tweet. We took into account also the ratio between the uppercase words and length of the tweet;
- *SentiWordNet.* We used the positive and negative scores to derive six features: positive/negative words count, the sum of the positive scores in the tweet, the sum of negative scores in the tweet, the balance (positive-negative) score of the tweet, and the standard deviation of SentiWN scores in the tweet;
- *Hu-Liu Lexicon*[5]. We derived three features from this lexicon: positive and negative words count, balance (# of positive words- # of negative words);
- *AFINN Lexicon*[6]. This lexicon contains two word lists labeled with polarity valences from -5 (negative) to +5 (positive). We derived 5 features from this lexicon: positive/negative word count, sum of positive and negative scores; overall balance of scores in the tweet;
- *Whissel Dictionary* [13]. This lexicon contains 8,700 Italian words with values of Activation, Imagery and Pleasantness related to each one. Range of scores go from 1 (most passive) to 3 (most active). We derived six features: average activation, imagery and pleasantness, and the standard deviation of the respective scores. We thought that an elevate score in one of these features may indicate an out-of-context word, thus indicating a possibly ironic comment;
- *Italian "Taboo Words".* Knowing the function of taboo words to trigger humour, catharsis, or to boost opinions, we decided to use a list of taboo Italian words that we extracted from Wiktionary[7];
- *Counter-Factuality* and *Temporal Compression* [9]. Frequency of terms that indicate an abrupt change in a narrative.

[3] http://sentiwordnet.isti.cnr.it

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[5] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[6] https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

2.1 Context: the 2014 Genoa Floodings

Due to the language of our system and the temporal proximity, we decided to test our system on a set of tweets related to the heavy rains and floodings that hit the city of Genoa on 9 and 10 October 2014. This event had a great impact on local media and partially on the Italian ones but it is not well known outside Italy, as it can be subsumed by the absence of a Wikipedia page about the event in English, but the event is covered in detail in the related Italian page[8]. The most critical events happened between 21:00 Oct. 9 when the Carpi river flooded Montoggio (a small town in the province of Genoa) and 01:00 Oct. 10 when various small rivers flooded the center of Genoa. In the aftermath, local authorities were criticized for bad crisis management and for being unable to foresee the event.

We extracted a total of 13,530 tweets, containing the keyword "Genova", between 19:00 Oct. 9 2014 and 16:00 Oct. 10 2014. We did not use geo-tagging to localise tweets because we found that only 754 tweets contained this information. In Figure 1 we show the number of tweets by hour during the critical period between the night of 9 October and the next day.

2.2 Bursts Detection

Apart from analysing the tweets using SA, we processed the flow of tweets with a simple topic burst detection technique based on a Poisson model for each time period [4]. The Poisson distribution, commonly used to express the probability of a given number of events occurring in a fixed interval of time. This technique allowed us to detect topic bursts to identify thematic changes in the flow of tweets. In [4], abnormally frequent events are detected when

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} > \epsilon \quad (1)$$

and $k > \lambda$. In our model, k is the number of occurrences of a given keyword X in one hour time interval, and λ is estimated by averaging the occurrences of X over the previous time periods. Therefore, if $Pr(X = k)$ is high, then the

[7] http://it.wiktionary.org

[8] http://it.wikipedia.org/wiki/Alluvione_di_Genova_del_9_e_10_ottobre_2014

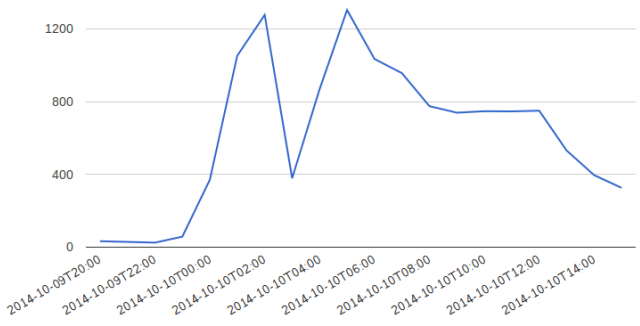

Figure 1: Number of tweets between 19:00 Oct. 9 2014 and 16:00 Oct. 10 2014 containing the word "Genova".

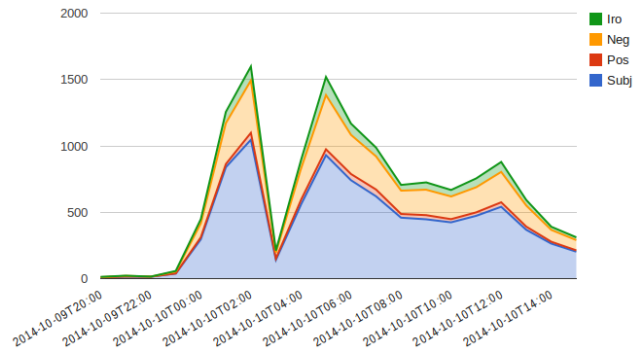

Figure 2: Number of tweets classified as subjective (*subj*), positive (*pos*), negative (*neg*) and ironic (*iro*), by hour. Note: *subj* includes the other classes.

frequency of the observed keyword is in line with previous observations, while a low value indicates that the frequency of the observed keyword is abnormally high (if $k > \lambda$). In our experiments, we set $\epsilon = 0$. We calculated bursts over hashtags, toponyms, and tweet fragments corresponding to Part-Of-Speech patterns that are usually more informative than a single word (such as adjective-noun, noun-adjective, noun-preposition-noun, verb-object etc...). Toponyms were detected using regular expressions filtered by means of a list of toponyms in the province of Genoa. Hashtags allow to describe tweet contents in a coarse way, while the textual fragments provide a fine-grained description. Toponyms are important clues to know where a specific event is occurring.

3. ANALYSIS

Within the selected time frame, IRADABE labelled 8922 tweets (65.9%) as *subjective*, 499 (3.7%) as *positive*, 3519 (26%) as *negative* and 1019 (7.5%) as *ironic*. IRADABE was trained on the complete Evalita SENTIPOLC dataset. In Figure 2 we show the number of classified tweets for each category, for each one-hour time period.

A quantitative evaluation of the obtained results is difficult since the tweets were not previously labelled with their correct labels (that is, no gold standard is available for these data). However, basing on the hypothesis that personal accounts are more probable to publish subjective tweets than organisational or news accounts, we were able to evaluate the correctness of the subjectivity classification. We manually classified the 50 most active accounts, as "personal accounts" or "news sources" (or aggregators). Globally, 59.2% of the tweets were estimated to come from news sources, and 40.8% from personal accounts. If we take into account the tweets labelled as *subjective*, the percentage increases to 70.5% from personal accounts and 29.5%, in line with the accuracy obtained by IRADABE at SENTIPOLC in subjectivity classification (67%).

We selected manually a set of hashtags, toponyms and textual fragments (we may call them *topics*) that we judged to be important from a disaster management perspective. For instance, we included the toponyms related to the affected zones (Montoggio, Bisagno, Sturla, Fereggiano,...) as they are listed on the Italian Wikipedia page on the disaster; we included hashtags like *#allertameteo* (meteo alert), *#protezionecivile* (civil protection agency), *#alluvionege* (Genoa flooding); we included fragments like "ondata di piena" ("surge wave"), "mancato allarme" ("missed alarm"), "invaso dal fango" ("flooded by mud"). Then we extracted for each time frame the list of trending hashtags, toponyms and topics, according to (1). We calculated accuracy as the number of trending important items (hashtags, toponyms or topics) divided by the total number of detected trending items, and coverage as the number of detected trending important items in all time frames divided by the total number of important items. The results of this evaluation are shown in Table 1. The results

Table 1: Average accuracy (*Acc*) and coverage (*Cov*) on relevant hashtags, toponyms and topics, calculated over all tweets (All) and over those labelled as subjective (Subj), positive, (Pos), negative (Neg) and ironic (Iro).

	Hashtags		Toponyms		Topics	
	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.
All	0.160	1.000	0.660	0.875	0.104	0.980
Subj	0.197	0.875	0.661	0.844	0.156	0.804
Neg	0.254	0.625	0.482	0.594	0.137	0.451
Pos	0.208	0.188	0.155	0.156	0.164	0.157
Iro	0.322	0.313	0.362	0.344	0.118	0.098

show that if we limit the analysis to subjective and negative tweets, we can obtain higher average accuracy (in other words, a higher percentage of the reported items for each time period are relevant) but at the expenses of coverage. Note that 15 and 16 of the items detected using the negative and subjective tweets, respectively, were not detected using the complete dataset. In Figure 3 we show the average accuracy, hour by hour, on all items, using the full dataset or only the subjective and negative tweets.

We manually analysed some of the tweets with an assigned polarity or irony label. We discovered that many positive tweets (58) were originated by a single account which was posting video links, adding "buona visione!" ("enjoy the show!") with an happy smiley, nevertheless their content (probably an automated posting). The second most frequent poster of positive tweets totalled only 8 tweets, all thanking the volunteers that helped in removing mud from the streets. Most positive tweets appeared to be encouragement messages to the population like "Forza Genova!" ("Come on Genoa!"). We can conclude that most of positive tweets are not pertinent from the disaster management per-

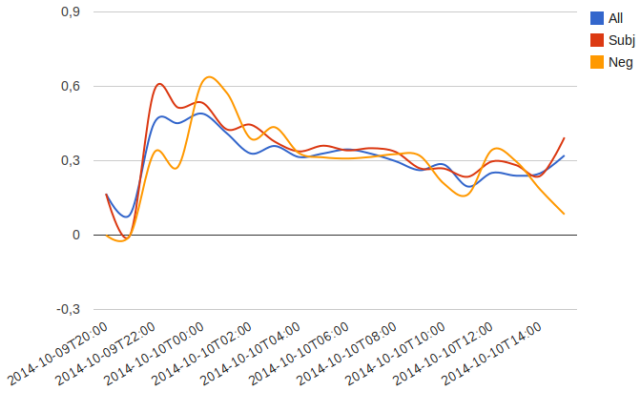

Figure 3: Hourly interpolated accuracy, averaged on all class of relevant items.

spective, corroborating our initial hypothesis on this class. We found a high rate of false positives among the tweets labelled as ironic, indicating that this task is yet a difficult one. However, among the correctly labelled ones, we were able to find many tweets that criticized the local authorities, such as "Ragazzi tranquilli #Renzi ha detto che non ci lascerà soli" ("Don't worry guys, #Renzi - the Italian prime minister - said that he won't leave us alone"). These tweets were probably detected because of the high number of politically themed ironic tweets in the SENTIPOLC training set. In the case of negative tweets, we were able to identify different kinds of negative feelings, from worry and fear to rage and frustration.

4. CONCLUSIONS

We applied a Sentiment Analysis system for Italian to a set of tweets covering the period of the Genoa floodings in October 2014, labelling tweets as subjective, positive, negative or ironic. We attempted to identify trending hashtags, topics and toponyms that may be relevant from a disaster management perspective in the different labelled subsets. Our system was able to identify subjective posts with good accuracy (around 70%). Our analysis shows that, although the use of the complete dataset provide a broader coverage of the ongoing event, focusing on a specific category of tweets may give insights on specific situations that may be missed when analysing the complete data. We found that positive tweets are not particularly useful in the analysis and they can be discarded, while a more fine-grained classification on the negative tweets may help to distinguish different kind of negative feelings, for instance fear. We plan to modify our SA system to improve its accuracy and take into account different types of feelings. We plan also to extend our experiments over a greater set of tweets around the event in order to improve the burst detection model.

Acknowledgments

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the progam "Investissements d'Avenir" (reference: ANR-10-LABX-0083). The National Council for Science and Technology (CONACyT-Mexico) has funded the research work of the second author (218109/313683 grant).

5. REFERENCES

[1] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[2] V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 50–57, Pisa, Italy, 2014.

[3] I. Hernández-Farias, D. Buscaldi, and B. Priego-Sanchez. IRADABE: dapting English Lexiconsto the Italian Sentiment Polarity Classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 75–80, Pisa, Italy, 2014.

[4] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 207–216. ACM, 2006.

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[6] R. Lee, S. Wakamiya, and K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.

[7] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

[8] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.

[9] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268, 2013.

[10] B. Robinson, R. Power, and M. Cameron. A Sensitive Twitter Earthquake Detector. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 999–1002, 2013.

[11] S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. *Proc. SemEval*, 2014.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[13] C. Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language 1, 2. *Psychological reports*, 105(2):509–521, 2009.