The Problem of Data Cleaning for Knowledge Extraction from Social Media

Emre Calisir[0000−0001−7505−7827] and Marco Brambilla[0000−0002−8753−2434]

Politecnico di Milano. Dipartimento di Elettronica, Informazione e Bioingegneria
Piazza Leonardo da Vinci, 32 – 20133 Milano, Italy
{firstname.lastname}@polimi.it

Abstract. Social media platforms let users share their opinions through textual or multimedia content. In many settings, this becomes a valuable source of knowledge that can be exploited for specific business objectives. In this work, we report on an implementation aiming at cleaning the data collected from social content, within specific domains or related to given topics of interest. Indeed, topic-based collection of social media content is performed through keyword-based search, which typically entails very noisy results. Therefore we propose a method for data cleaning and removal of off-topic content based on supervised machine learning techniques, i.e. classification, over data collected from social media platforms based on keywords regarding a specific topic . We define a general method for this and then we validate it through an experiment of data extraction from Twitter, with respect to a set of famous cultural institutions in Italy, including theaters, museums, and other venues. For this case, we collaborated with domain experts to label the dataset, and then we evaluated and compared the performance of classifiers that are trained with different feature extraction strategies.

Keywords: Social media · Knowledge discovery · Data cleaning · Data Wrangling · Text classification.

1 Introduction

Social media has become one of the most powerful information channels in the digital age. Today, more than 1.6 billion social network users actively create content on these platforms for more than two hours each day.[1] For instance, in Twitter, the well-known social media platform where the users are writing short texts called ”tweets”, the creation of new content is so fast that 100 million Twitter users post 500 million tweets every day.[2] Consequently, the immense amount of user generated data provides a good opportunity for every field of study.

[1] Statista, on social media usage. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/, Last accessed 4 Apr 2018.

[2] Omnicore Agency, https://www.omnicoreagency.com/twitter-statistics/, Last accessed 4 Apr 2018.

On the other hand, the human-generated content brings many challenges for extracting knowledge for specific topics. The traditional rule-based content extraction features provided by social network platforms (e.g., through API) are based on keyword or metadata search queries. In principle, these systems are designed to retrieve topic-specific data. However, in practice they are not able to address the problem properly, especially for what concerns the relevance of the collected content: indeed, the risk is that they return related and unrelated content altogether, due to various issues, including synonyms, shared keywords across topics, and so on. This is also due to the noise in the data, specifically caused by grammar and spelling errors, or multiple meanings of words [1]. Consequently, it is strongly needed to build new systems that are capable of obtaining clean and trusted topic-specific datasets from social media, after a first data collection phase (typically built on keyword-based search) is performed.

In this study, we propose to use supervised learning, with its ability to learn significant features of topic-relevant posts from a trusted labeled dataset, by devising an approach that is applicable to all kinds of social media platforms. We define our problem setting and research questions as follows:

Input: Human-generated textual content shared on social media platforms; and definition of a specific topic, named entity or context of interest.

Collected data: Set of human-generated textual content collected by querying the social media platform APIs based on the topic, entity or context of interest. Content collected in such way may include irrelevant data, due to synonyms, common keywords, excessively broad context, and so on.

Research Question: Given a set of social media content items collected as above, extract a sub-selection of content items if and only if they are actually relevant to the topic or context of interest.

In the paper we propose a general approach to this issue and then we test the approach over a real use case on Twitter data.

The paper is organized as follows. In Section 2, we provide prominent studies in knowledge extraction from social media data. Then in Section 3, we explain our methodology by presenting our applied machine learning algorithm and the different feature extraction strategies including n-grams, word2vec, the combination of additional features with word2vec and the dimensionality reduction. In Section 4, we explain our experiment on a real-world use case, which is the exploration of tweets about cultural institutions of Italy. In Section 5, we show how the different classifiers having different feature extraction strategies impact the accuracy of Machine learning classifier. And finally, in Section 6, we conclude our study and give information about our future work.

2 Related Work

Researchers have shown great interest to obtain a clean dataset from social media data for several years. In one of the earliest studies [2], the authors created an early earthquake alarm system based on tweets in order to deliver the announcements much faster than Japan Meteorological Agency. In that study, the researchers trained a Support Vector Machine (SVM) classifier by extracting a variety of features such as keywords, the number of words, and the context of target-event words. In another study [3], the purpose is to detect influenza-like illnesses (ILI) from tweets. It is indicated that Bag-of-Words (BoW) based Logistic regression model could achieve a correlation of .78 with statistics of Centers for Control and Prevention. In another health-related study [4], the authors filtered out the non-relevant content for a health topic related study, and they implemented a binary logistic regression model with unigram, bigram and trigram word features. In a recent health-related study [5], the authors investigated the prevalence and patterns of abuse of specific medications based on an automatic supervised classifier trained by annotated tweets. In [6], the authors tracked baseball and fashion topics over streaming tweets by implementing unigram language models that are smoothed using a normalized extension of stupid back-off. In [7], the authors argue that BoW classifiers fail to achieve good accuracy in short texts. They propose to use Multinomial Naive-Bayes and a collocation feature selection algorithm to increase the performance of BoW. In [8], the authors built a classifier to filter out noisy events for an event detection system from the Twitter stream. In another study [9], the authors did not filter out the non-relevant tweets from a chosen topic, but they performed a more general analysis on all tweets containing trend topic hash-tags to assess whether they are credible in terms of relevancy. They evaluated the performance of different algorithms such as SVM, decision trees, decision rules, and Bayes networks, and they achieved the best performance with a J48 decision tree method.

On the other hand, in some studies, the researchers have needed to enrich the tweet information with external data sources such as search engines [10], page content of linked embedded URLs in tweets [11], and other data sources such as Wikipedia [12]. These approaches are targeting to increase the amount of text data for tweets, but they don't enable to build real-time information filtering systems [13].

In addition to supervised learning based studies, in a recent research [14], the researchers focused on an unsupervised learning method, which is based on a pooling method combining both Information retrieval (IR) and Latent Dirichlet Allocation (LDA) in order to prune the irrelevant tweets.

As it can be understood from all of the studies explained in this section, there is a continuous interest in topic-based information filtering on Twitter. With the approach of comparing n-grams and word embedding techniques and having a real-world use case, we believe that our research could contribute to the existing studies.

3 Methodology

The most basic solution to knowledge extraction about a topic using social media data is to build a rule-based system. In this approach, there are explicitly defined rules, such as storing all of the records containing specific keywords, hash-tags or account names. However, the disadvantage of this approach is that it produces very noisy data. To prevent this issue, we propose to implement a machine learning system, with a classifier trained on the specific context. Consequently, the system could recognize the relevant and non-relevant tweets; and it becomes possible to obtain a clean dataset in given context. Figure 1 represents the high-level flow diagram of the proposed method.

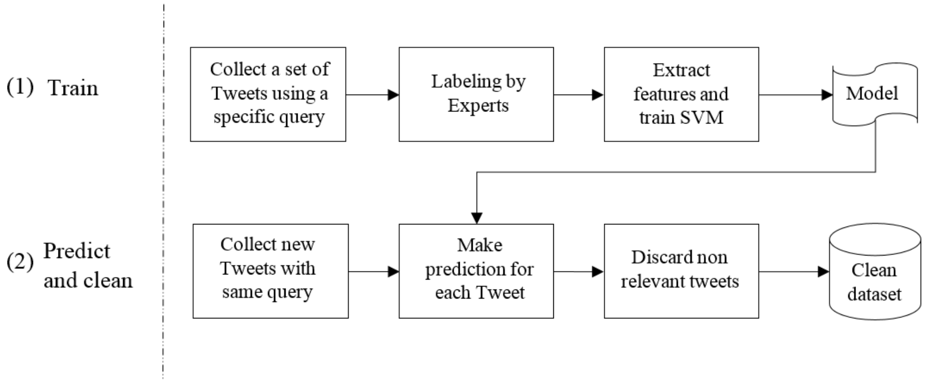
(1) Train
Collect a set of Tweets using a specific query
Labeling by Experts
Extract features and train SVM
Model
(2) Predict and clean
Collect new Tweets with same query
Make prediction for each Tweet
Discard non relevant tweets
Clean dataset

Fig. 1. Proposed Data Cleaning Method for Knowledge Extraction

3.1 Machine Learning Classifier

In our study, we used Support Vector Machine (SVM) [15] with Linear Kernel which is the recommended algorithm for text classification on high dimensional sparse data [16].

3.2 Feature Extraction Strategies

In terms of building the classifier, there are several approaches such as using only the text content [17], or combining the text content with additional specific features [18]. In this study, we built four different models to observe how the different feature sets impact the accuracy of the SVM-based classifier.

Table 1. Attributes of Tweet: Each attribute points out the relevancy of tweet.

Attribute	**Data type**	**Context**
Text	Text value	Tweet
Count of favorites	Numerical value	Tweet
Count of retweets	Numerical value	Tweet
Count of lists	Numerical value	Author
Count of tweets	Numerical value	Author
Count of accounts followed by	Numerical value	Author
Count of accounts followed	Numerical value	Author
Is geographic enabled	Categorical value	Author
Is verified account	Categorical value	Author
Is default profile	Categorical value	Author
Source of posts	Categorical value	Author

Model 1 (only n-grams): N-grams is one of most widely used text representation techniques. This method combines information derived from n-grams (consecutive sequences of n characters or n words) with a simple vector-space technique [19].

Model 2 (only word2vec): Basically, word2vec creates semantic connections between words. It produces word vectors with deep learning via word2vec's skip-gram and CBOW models, using either hierarchical softmax or negative sampling [20]. This technique enables positioning word vectors in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

Model 3 (word2vec+additional features): In the previous feature extraction strategies, we were considering only the text data. However, there are useful attributes that could be added into the same machine learning pipeline. In this model, we used all attributes shown in Table 1 as input features.

Model 4 (PCA applied on word2vec+additional features): In this model, we apply Principal Component Analysis (PCA) over the features of Model 3 in order to reduce the number of features. PCA transforms existing features to new features, which better represent the data, with fewer features having higher variance. Another advantage of using PCA is that it is one of the methods to prevent overfitting of the model.

4 Experiments

In our real-world scenario, the purpose is to obtain a clean dataset from Twitter about cultural institutions in Italy, including theaters, museums, and other venues. We extract possibly relevant tweets by querying the social network using the most typical keywords and ashtags referring to those venues, and then we aim at cleaning the collected data by removing all the non-pertinent contents.

In our experiment, we focused specifically on tweets related with Pompei, Colosseum and Teatro Alla Scala. Pompei is a city in southern Italy's Campania region overlooked by the active volcano at Vesuvius. The Colosseum is the oval amphitheater in the center of the city of Rome. Built of travertine, tuff, and brick-faced concrete, it is the largest amphitheater ever built. And finally, Teatro alla Scala is a famous opera house in Milan.

We applied our method in this use case in the following way: (1) We collected tweets using Twitter Search API. (2) We provided the tweets to the experts for labeling process. (3) We received a set of 726 tweets from the experts, in which the non-relevant and relevant tweets are equally distributed. (363 non-relevant and 363 relevant tweets) (4) We built four different Machine Learning models having the different type of features as described in Section 4.2. (5) We evaluated the prediction performance of the models using cross-validation.

4.1 Data Collection Phase

Twitter enables accessing to the publicly shared posts with Search and Stream API. The standard version of these services gives random 1% of tweets in a given criteria. For our real-world experiment, we used Search API to collect tweets posted in a specific time period. In order to determine the scope of the use case, we worked with subject matter of experts of famous cultural institutions of Italy, and we finally prepared the following search query:

"@teatroallascala or #TeatroallaScala or Colosseo or #colosseo or Pompei or #Pompei"

Also, we specified the date period of tweets as:

"since:2017-12-01 until:2018-01-31" .

Tweets contain a variety of fields that could be useful for a classification task. For our use case, we have decided to use the attributes described in Table 1.

4.2 Annotating the Relevant and Non Relevant Tweets

In terms of giving correct decisions in labeling process, we collaborated with the experts. By looking at the textual content of the tweets, the experts manually labeled a set of tweets as relevant and non-relevant, depending on whether they are in context. In this process, they did not eliminate the tweets regarding its written language due to the limited size of our dataset.

Below, there is an example of a relevant and a non-relevant tweet. It is clear that the non-relevant tweet should be eliminated from dataset since it is a commercial advertisement for an hotel in Pompei. In contrast, the relevant tweet contains valuable information about the historical background of Pompei, and it should remain in the dataset.

Table 2. Word Similarities in trained word2vec Model: for each sample word, the list of the top-3 most similar words are shown, with the respective similarity score.

Word	First similar word	Second similar word	Third similar word
colosseum	rome (0.994)	roma (0.994)	coliseum (0.994)
colosseo	anfiteatro (0.995)	travel (0.994)	italia (0.994)
scala	aux (0.993)	camelias (0.992)	milano (0.992)
pompei	retweeted (0.988)	nuovi (0.979)	settembre (0.978)
roma	rome (0.995)	metro (0.994)	colosseum (0.994)
italia	anfiteatro (0.995)	rome (0.995)	colosseo (0.994)
italy	travel (0.998)	davanti (0.997)	photography (0.997)

Non-relevant Tweet: Best #Hotel Deals in #Pompei #HotelDegliAmiciPompei starting at EUR99.60 https://t.co/5DxkKn4o69 https://t.co/akyJoBLwq3

Relevant Tweet: Pompei Hero Pliny the Elder May Have Been Found 2000 Years Later https://t.co/PyR2rP1Xpe #2017Rewind #archaeology #archeology #history #Pompei #rome #RomanEmpire #history

4.3 Feature Transformations

For this use-case, by using the feature extraction strategies explained in Section 4.2., we transformed the tweets into a convenient structure for the classifier in the following ways.

Model 1 (only n-grams): As an input of this model, we initially transformed the text of tweet to the word sets of unigrams, bigrams, and trigrams as a similar approach to [4]. In addition, we applied Term Frequency - Inverse Document Frequency (TF-IDF) technique over the n-grams.

Model 2 (only word2vec): For the word2vec based models (Model 2, 3, and 4), we built the vocabulary using the full-text content of 2558 tweets. The word2vec vocabulary size is equal to 34029. We determined to use 25-dimensional vectors due to the fact that we have a limited word2vec vocabulary. Indeed, after doing several experiments on different vector dimensions, we observed that the performance of word2vec models having higher vector dimensions are less successful in constructing semantic connections when the vocabulary size is low.

The word similarities of our word2vec model is shown with a small example in Table 2. It is not very surprising that most similar words of *scala* are *aux*, *camelias* and *milano*, because Teatro Alla Scala is the opera house located in Milan, where the famous ballet *La Dame aux camélias* is staged. In contrast, we observe that the limited vocabulary size caused some unexpected similarities as well, as in the case of *pompei* and *retweeted*.

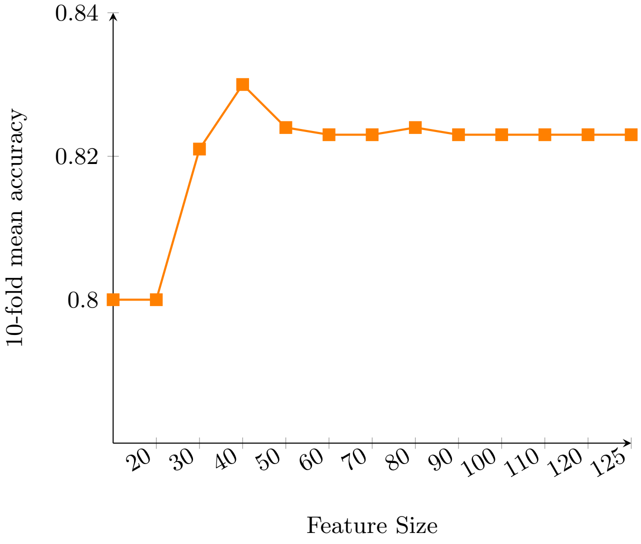

Fig. 2. Principal Component Analysis: The change in the accuracy based on the target feature size

Model 3 (word2vec+additional features): In Table 1, we present the additional features that describe different aspects of tweets. In this model, we are combining word2vec representation of text content with the categorical and numerical features. In order to use them efficiently in a classifier, we apply Min-Max scaling to numerical variables and One-hot encoding to categorical variables. Consequently, we obtained more than 125 features.

Model 4 (PCA applied on word2vec+additional features): In this model, we applied PCA over the feature set of Model 3 due to the fact that one-hot encoding increased the feature dimensions in a significant manner. Here, we determined the size of PCA applied feature set by measuring the accuracy of classifier for each possible dimension. The model reaches to the best accuracy when the dimension size is 40 (see Figure 2).

5 Results

By using the equally distributed dataset described in Section 5, a rule-based system which filters tweets regarding the existence of specific hash-tags or keywords could make a prediction with an accuracy of 0.5. However, even though we have a very limited dataset and word2vec is unable to perform perfectly with a small amount of data, all of the classifiers achieved similar and high prediction scores as shown in Table 3. Here we determined to use 10-fold cross validation due to the fact that our dataset is not very large, and cross-validation could give more trusted results.

Table 3. 10-fold mean average values of models for each feature extraction strategy

Machine Learning model	Accuracy	Precision	Recall	F1-Score
Model 1	0.84	0.844	0.832	0.838
Model 2	0.816	0.785	0.869	0.824
Model 3	0.823	0.834	0.807	0.819
Model 4	0.83	0.844	0.81	0.826

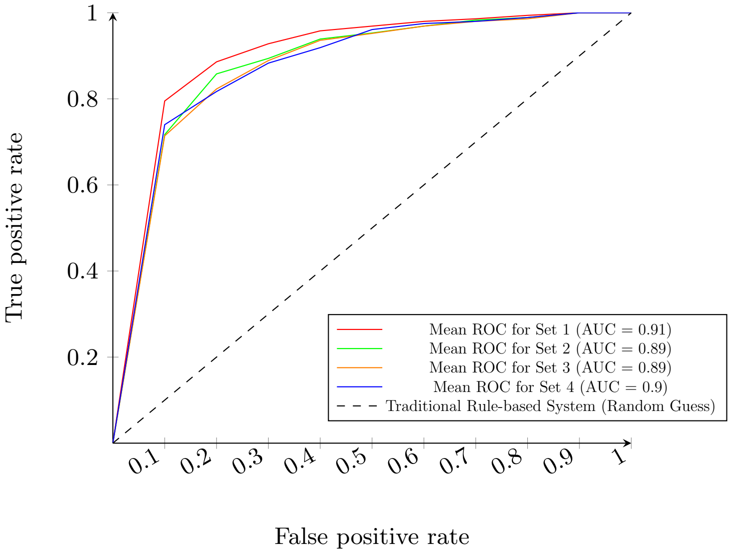

Fig. 3. Comparison of ROC graphs: for each model, the ROC curve and AUC values are shown.

In addition to the performance indicators shown in Table 3, it is also important to interpret a classifier with its structure of Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) scores. The ROC curves clearly illustrate the prediction success of the models (see Figure 3).

When we compare the models between each other, it is easy to see that the n-grams based model (Model 1) had the best performance among all models. One possible explanation could be the limited word2vec vocabulary. Actually, word2vec requires very large amounts of training data to provide better results. On the other hand, we observe that combining multiple features had a positive impact on classifier performance. Also, the dimensionality reduction technique increased the accuracy as expected.

6 Conclusions and Future Work

In this study, we proposed Machine learning based methods to obtain a clean and trusted topic-specific dataset from Twitter. In a real-world use case, we proved that our approach achieves high accuracy even though the training dataset is very limited. In the future studies, we will collect more tweets to build a larger corpus for word2vec models, and also increase the size of annotated training dataset. Also, we will explore the impact of adding additional tweet-specific features that could improve the performance of our models, such as number/presence of hashtags, hyperlinks, user mentions, and length of the tweet. Moreover, we will analyze the performance of classifiers by using separate tweet data sets for each language, and we will observe how n-grams and word2vec perform on the uni-language corpus.

Acknowledgements

We gratefully acknowledge Fluxedo S.r.l.[3] for sharing the data used for our experiments and the Osservatorio MIP Innovazione Digitale nei Beni e Attività Culturali[4] for the useful insights and discussion on the matter.

References

1. Salloum, SA, Al-Emran, M, Monem, AA, Shaalan, K: A survey of text mining in social media Facebook and Twitter perspectives. In: Advances in science, technology and engineering systems journal (2017).
2. Sakaki T, Okazaki M, Matsuo Y: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, pp. 851–860, New York, USA (2010).
3. Culotta A: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the first workshop on social media analytics, Washington, D.C., pp. 115–122 (2010).
4. Paul M.J, Dredze M: Discovering health topics in social media using models. In: PLoS ONE 9, e103408 (2014).
5. Sarker, A, O'Connor, K, Ginn R, Scotch M, Smith K, Malone D, Gonzalez, G: Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. In: Drug Safety, 39(3), pp. 231–240 (2016).
6. Lin J, Snow R, Morgan W: Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In: KDD, pp. 422–429 (2011).
7. Khan MAH, Iwai M, Sezaki K: An improved classification strategy for filtering relevant tweets using bag-of-word classifiers. In: Journal of information processing, 21(3), pp. 507–516 (2013).
8. Kunneman F, Bosch A: Event detection in Twitter: A machine-learning approach based on term pivoting. In: BNAIC, pp. 65–72 (2014).

[3] http://www.fluxedo.com/

[4] https://www.osservatori.net/it_it/osservatori/innovazione-digitale-nei-beni-e-attivita-culturali

9. Castillo, C, Mendoza M, Poblete B: Information credibility on Twitter. In: Proceedings of the 20th international conference on World Wide Web. ACM, Hyderabad, USA, pp.675–684 (2011).
10. Bollegala, D, Matsuo Y, Ishizuka M: Measuring semantic similarity between words using Web search engines. In: Proceedings of the 16th international conference on World Wide Web (WWW2007) ACM Press, New York, pp. 757–766 (2007).
11. Yang S, Kolcz A, Schlaikjer A, Gupta P: Large-scale high-precision topic modeling on Twitter. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '14). ACM, New York, pp. 1907–1916 (2014).
12. Banerjee S, Ramanathan K, Gupta A: Clustering short texts using Wikipedia. In: SIGIR, pp. 787–788 (2007).
13. . Li Q, Liu X, Shah S, Nourbakhsh A: Tweet Topic Classification Using Distributed Language Representations. In: Proceedings of the 2016 IEEE/WIC/ACM international conference on Web intelligence. Nebraska, USA (2016).
14. Hajjem M, Latiri C: Combining IR and LDA topic modeling for filtering microblogs. In: Procedia computer science, 112, pp. 761-770 (2017).
15. Joachims T: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the European Conference on Machine Learning, pp. 137–142, Springer, Berlin (1998).
16. Lewis JP: Tutorial on SVM, In: CGIT Lab, USC (2004).
17. Sun A: Short text classification using very few words. In: SIGIR, ACM, pp. 1145–1146 (2012).
18. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M:Short text classification in Twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval (2010).
19. Damashek M: Gauging similarity with n-grams: Language independent categorization of text. In: Science, 267(5199), pp. 843–848 (1995).
20. Mikolov T, Chen K, Corrado G, Dean J: Efficient estimation of word representations in vector space. In: Proceedings of workshop at ICLR (2013).