Automated Detection of Urban Flooding from News

F. Zarei[a], and M. Nik-Bakht[a]

[a]Department of Building, Environment and Civil Engineering, Concordia University, Canada
E-mail: f_zaire@encs.concordia.ca, mazdak.nikbakht@concordia.ca

Abstract –
Although minor overflows do not cause a huge amount of loss; as the number of such overflows increases, the amount of water wasted, and the compound consequent challenges become considerable. Therefore, detecting overflows, investigating their cause root and resolving them in a timely manner are among new needs for infrastructure managers. This paper suggests a new method for detecting distributed water overflows by extracting the flood information (such as the date and location of the incident) from the news. As a case study, we crawled Montreal newspaper and news websites to detect the related news to urban flooding and their detailed information. We trained several classifiers to identify news relevant to flood. Our experiments illustrate that by applying mutual information method for feature selection and employing support vector machine (SVM) architecture as the classifier, relevant news can be detected with an accuracy (F-measure) of above 80%. Such actionable information can help infrastructure managers with a wide range of decisions from repair and maintenance of existing systems, to capacity evaluation for new designs.

Keywords –
Classification; Montreal newspaper; Urban Flooding; Water overflows

1 Introduction and background

Nowadays, anthropogenic climate change alters weather patterns with significant shifts in climate normal and extremes [1]. On the one hand, having more extreme events such as heavy rainfalls could lead to an increase in the risk of urban flooding. On the other hand, the aging infrastructure, encroachment on drainage canals, and reduced natural drainage introduce additional risk factors. Heavier rainfalls in aging infrastructure may generate urban flooding [2]. This is particularly increasing the frequency and magnitude of distributed incidence which is also known as compound effect.

At the same time, 'responsible citizenship' is an emerging phenomenon with several benefits such as citizen engagement, detecting public service performance and citizens 'priorities via feedbacks [3]. Reporting the civic complains in the media in recent years is supporting in this regard. In fact, it is a belief that citizens are inherently capable to solve their own problems at a higher level, collectively and collaboratively.

The citizens are enabled by resolving the urban issues to monitor, record and report the social problems together [4]. Several social media platforms such as Twitter, Instagram and complaint boards provide channels for the citizens to get engaged in different levels of problem solving and decision making. Using these platforms, citizens can take a photo from a civic phenomenon, add a comment, geotag, and share it. In fact, it is argued that citizens gain more awareness by the aid of these platforms and they can help the authorities to learn about the issues faster and more efficient. In [5], the authors use Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) for detecting traffic accidents from social media data, specifically Twitter, and compared their results with the data of 15,000 loop detectors. It is concluded in this research that around 66% of the tweets which are relevant to the accidents are consistent with the facts about the accidents.

By using tweets, the authorities can gain a better understanding of the number of involved citizens with the issues and the intensity of problems [6]. In this regard, after the emergence of smart cities, it is attempted to apply new technologies and ideas to improve social, political and economic systems. It is aimed that such technologies could help the citizens to increase their level of participation in civic issues.

In [7], the authors provided a platform named "Citicafe". This platform has an exchangeable based interface to improve the citizen engagement by allowing direct messaging. It can help citizens to gather information related to their neighborhoods and report their problems. As another example, Indian government authorities collaborated with Twitter to launch "Twitter Seva" [8] whose goal is the enhancement of public service engagement by providing a platform for following complaints in a real-time manner. In this platform, the citizen interactions are limited to one whose communicated since there are no chat interfaces.

Considering board range of events reported in social media, one of the high-level goals of our research work is to automatically detect the overflow incidents from people's inputs and also to communicate warning alarms. We need to study the source of any observed and reported on-the-ground water to know whether it is related to flood, overflows or pipe leakages.

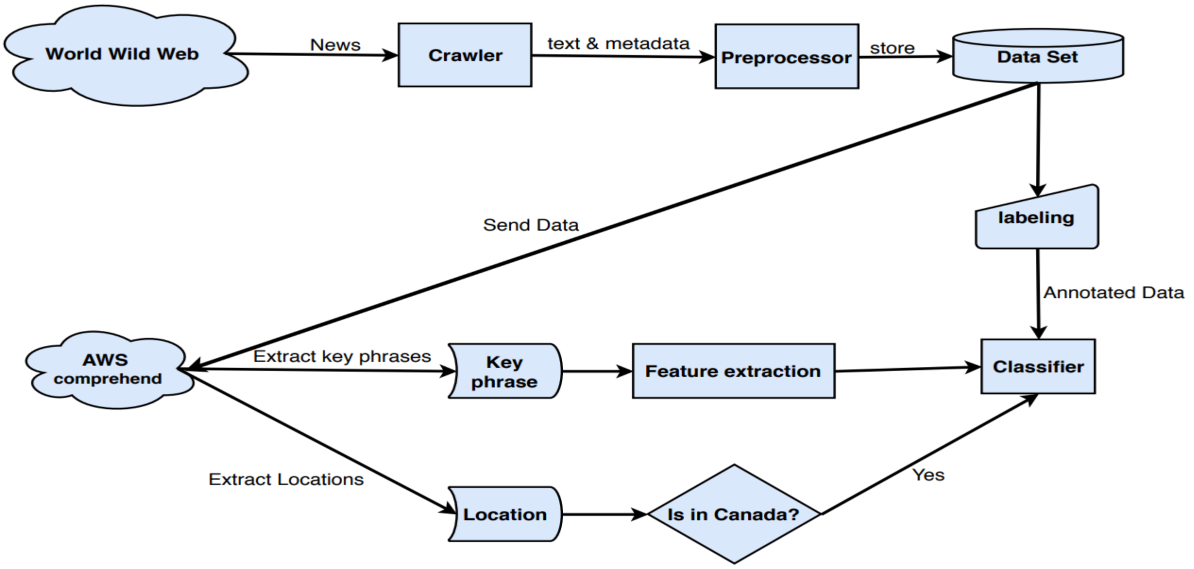


Figure 1. The architecture of the proposed method

In this regards, a binary classifier is required to categorize floods in a more specific context; from a nice rainy day to some serious failures. In the ideal case, not only we should be able to automatically detect flood spots in the city, but also we must detect the source and type of flood. In fact, we must be able to realize whether, with a high chance, it is a natural phenomenon or related to an infrastructure shortcoming.

With this aim, we collected and annotated news, and went through their content to find more related words. Considering co-occurrence of terms, we need to exclude the news which are not about water but include the terms which are normally used in discussing articles. On the first step, by searching the term "flood" on flood newspaper websites, we let the website's search engine to screen the news. Then, the obtained news was passed to Comprehend, which is Software as a Service (SaaS) provided by the amazon web services (AWS) [9]. Comprehend helped us to extract entities and key phrases from the news sentences. On the next step and after generating the relevant features manually, the required features for classification were selected by applying a mutual information method [10]. Having the features, Bayesian classifier (Naïve Bayes) [11], Meta algorithm (AdaBoost) [12], Decision tree (Id3) [13], Neural network (Multi-layer perceptron (MLP)) [14], Kernel estimation (KNN) [15] and Support vector machine (SVM) [16] classifiers were tested to label the news as relevant and irrelevant to the flood. Finally, the accuracy of the classifiers was evaluated in the sense of F-measures.

The remainder of this paper is organized as follows. The overall process flowchart of the proposed method is discussed in Section 2. Section 3 provides a discussion about the approach of crawling and also the data collection process. Applied feature extraction and feature selection approaches are explained in Section 4. In Section 5 the method of classification is provided. Section 6 presents the simulation and the obtained results, and finally, concluding remarks are provided at the end of the paper.

2 Proposed method

In this section, the entire flowchart of the proposed method will be discussed. As mentioned before, the main purpose of this work is detecting flood events in city of Montreal. In this regard, we concentrate on the flood-related news reported in Montreal newspapers. After gathering such news, the topic of each news article, which is a sentence, is considered as the input of the system. By using the Application Programming Interface (API) of the AWS, the existing name-entities and the key-phrases of the sentence are extracted. The news is considered irrelevant to our research if its extracted location is outside Canada. In such cases, the system goes to the next sentence. Otherwise, each word of the news article is searched in the provided word-list. If that word exists in the *i-th* place of the word-list, the *i-th* element of the feature vector becomes 1 and remains zero otherwise. Next, the feature generator provides an appropriate matrix of data with their features for the classifier. Finally, the classifier decides whether the news is related to the flood. This procedure is represented in Figure 1.

3 Data extraction

3.1 Data collection

The applied dataset was selected from the most popular Montreal local English newspaper, “Montreal Gazette” [17]. In this regard, with the aid of search engine of this website, we filtered the news archive and selected the ones which contain the word “flood”. On the next step and after such filtering, all the related news gathered into our database by building a crawler for this website in python.

3.2 Corpus construction

Once we crawled all the news form the mentioned newspaper, we had to reorganize the collected data in a format which is easy to use for our research purposes. The data of the gathered collection consist of 6 attributes including title, the body of the news, the category of the news, the date, the link, and the link of images.

3.3 Data annotation

In the next step, we assign a number to each sentence manually, which indicates the relevance of the news article. The news was divided into two categories named as relevant and irrelevant. The relevant news examples to the flood were labeled as ‘+1’ and the irrelevant ones labeled as ‘-1’. “Leaky pipe floods Van Horne Ave., creates Sunday traffic woes” and “More than a year after floods, family of 7 will soon be homeless” are two examples for each group, respectively.

4 Feature extraction and feature selection

Feature extraction and feature selection are two important processes which should be done prior to classification to improve its accuracy and decrease its computational complexity. Therefore, feature selection has a considerable importance in the scope of this paper. The methodology of feature selection applied in this work can be explained in two main steps: keywords extraction, and filtering.

4.1 Keyword extraction

The main aim of this step is extracting the ‘name entities’ such as the locations and extracting key-phrases of each news article. The location of a news is important in this research since our goal is finding the place of flooding. We applied AWS Comprehend service to extract the locations mentioned in the news as well as their key phrases. Amazon Comprehend is a natural language processing (NLP) service that uses deep learning to find the meaning and insights in texts. The topic of news was the input of the Comprehend service and the output was the location and key phrases of the news. We consider these keywords as the features for the classification. However, not all the extracted keywords can be beneficial for the classification. Firstly, the features applied in classification must be able to discriminate the data. Moreover, some of the extracted features are redundant and can deteriorate the classifier. Therefore, the features need to be filtered prior to the classification. The way of filtering of the features is described in the next section.

4.2 Manual selection of related and unrelated key-phrases

In this step, we calculate the occurrence of each feature in our dataset. Then, those features which are not very common in our corpus are deleted from the list of features. These features are the ones whose occurrence is less than 5 times in sentences. After this reduction, we still need to reduce the features to the ones related to flooding. Hence, we were through the list and deleted the irrelevant features. Finally, an automatic feature selection method was applied to the list of remaining features to make the features ready for classification.

4.3 Automatic feature selection by applying mutual information

There are many different approaches for feature selection. In this work, we focus on the information-theory-based approach. This method has attracted the most attention since it can detect nonlinear correlations among the features. Mutual information is an important concept in information theory. It can evaluate the relevance between two random variables X and Y. In this paper, the mutual information of two variables $word\ (w)$ and $topic\ (t)$ is defined as below:

$$\boldsymbol{MI(w,t)} = \mathbf{log}\big(\boldsymbol{P(w|t)}\big) - \mathbf{log}\big(\boldsymbol{P(w)}\big). \tag{1}$$

Here,

$$P(w|t) = \frac{\mathrm{NDTW}}{Total\ number\ of\ Docs\ with\ the\ topic\ t}$$

$$P(w) = \frac{NDW}{Total\, number\, of\, Docs}$$

Here, NDTW stands for the number of documents with the topic t containing the word w and NDW stands for the number of documents containing the word w. We apply this method to select a subset of highly discriminant features. Hence, the features which have the capability of discriminating the data of different classes are selected.

5 Classification

In a typical supervised classification problem, the main goal is to train a classifier by using a dataset $U = \{(X^1, t^1), \ldots, (X^n, t^n)\}$ of n labeled instances where each instance x^i is characterized by d features, i.e. $X = (X_1, \ldots, X_d)$, and a label which indicates the class that it belongs to.

In this section, the main goal is to classify the news into 'related to the flood' and 'nonrelated to the flood'. Different applied classification methods are Naïve-Bayes classifier, AdaBoost, Id3, MLP and SVM. In the next section, the results of these classifiers are compared.

In order to evaluate the classifiers diagnostic tests were performed and three metrics were considered: precision, recall and F-measure.

$$precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F\_measure = 2.\frac{precision.\, recall}{precision + recall}$$

Here, *tp* is the number of data samples with the real label of positive which are labeled as positive by the machine. Furthermore, *tn* is the number of data samples with the real label of negative which are labeled as negative by the classifier. Similarly, *fp* is the number of data samples with the real label of negative which are labeled as positive by the machine. Finally, *fn* is the number of data samples with the real label of positive which are labeled as negative by the classifier.

6 Result and Evaluation

After search in "Montreal Gazette" newspaper data, the term "flood", 957 news items were found, started from September 16, 2009. This extracted news was manually labeled: 528 of them are labeled as relevant and 429 as irreverent.

By applying the feature extraction through Comprehend service and omitting the irrelevant features, the number of selected features for the classification was reduced to 98. In the next step and by applying the mutual information method, the number of features reduced to 50.

These features were passed to the classification agents whose results are reported in Table 1. For the evaluation, 10-fold cross validation was applied and precision, recall, and F-measure were calculated.

Table 1. Results of the classifiers

Classifiers	Precision	Recall	F-measure
Naïve Bayes	0.741	0.887	0.807
AdaBoost	0.711	0.932	0.805
Id3	0.707	0.887	0.787
MLP	0.697	0.870	0.774
KNN	0.751	0.859	0.801
SVM	0.762	0.832	0.824

The importance of precision and recall depends on the application. In some cases, having a higher precision is critical and in some other cases the recall has such importance. However, generally it is desired to consider both of them and evaluate the F-measure. Comparing the accuracy of the classifiers, it is seen that the SVM classifier classified the data more accurately. This is due to the characteristic of the SVM classifier that makes it a good choice for providing an adequate discrimination boundary for double class classification problems.

7 Demonstration of the results

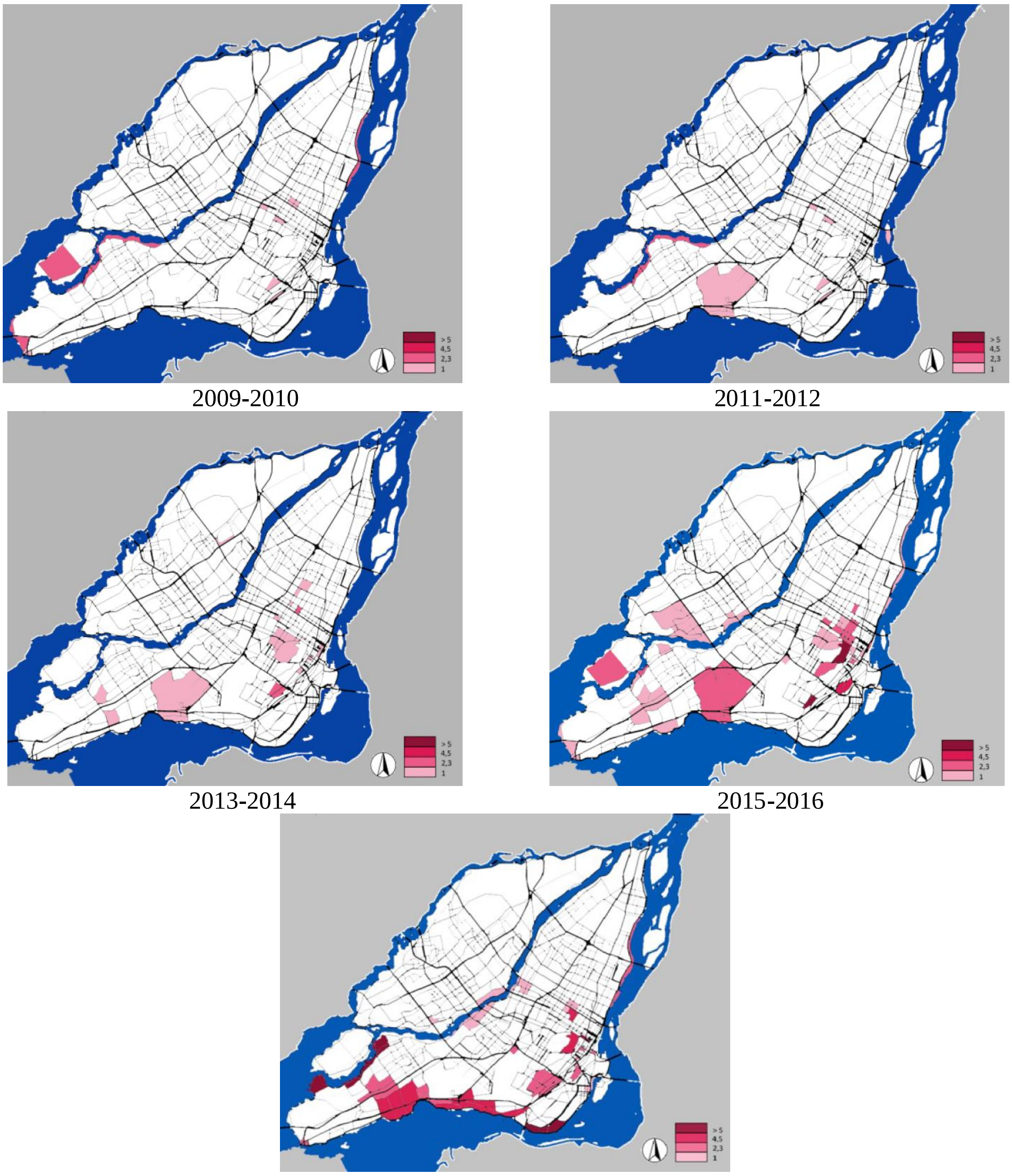
After extracting the locations and finding the relevant news, results were encoded as colors and were visualized. Figure 2 presents number of news mentioned flooding in "Montreal Gazette" from 2009 to 2018. As seen, the number of sections in which flood is reported is increasing through the time. Also, the colors are getting more saturated which indicates the higher number of flood report in that area.

8 Conclusion

Detecting the small water overflows and finding their resources are important especially when the cause of water overflows is a malfunction of the infrastructure. In this work, we attempted to detect news about the flood as either a natural or a man-made phenomenon, reported in Montreal newspaper. In this regard, 4 classifiers were trained and the results were compared. In the end, we reached more than 80% accuracy in terms of the F-measure score, which belongs to the SVM classifier by applying mutual information method for feature selection. The results of this research are useful for infrastructure managers who make decision about the maintenance of different infrastructure.

One of the main limitations of this work is that our corpus is quite small for the learning process since we just applied the text of one newspaper. Hence, more data sets should be applied to improve the performance of this work.

Following this work, we can use more properties from the website contents such as images of the news. In this work, we just classified the topic of the news. However, most of the news in the newspapers have images. Hence, by applying some image processing techniques, we can understand whether the news is related to the flood or not. By combing this results with the classifier results, we can increase the accuracy of the classification.



2009-2010

2011-2012

2013-2014

2015-2016

2017-2018

Figure 2. Number of flood reported in “Montreal Gazette” in different parts of Montreal

References

[1] IPCC (2013) Climate Change 2013: *The Physical Science Basis*. Stocker T, Qin D (Co-Chairs) Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Working Group I Technical Support Unit, Switzerland.

[2] Markolf, S. A.; Hoehne, C.; Fraser, A.; Chester, M. V. & Underwood, B. S. Transportation resilience to climate change and extreme weather events – Beyond risk and robustness Transport Policy, 2019, 74, 174 - 186

[3] M. M. Skoric, Q. Zhu, D. Goh and N. Pang. 2015. Social media and citizen engagement: A meta-analytic review. New Media and Society, Sage Publishing.

[4] S. K. Prasad, R. Patil, S. Beldare, and A. Shinde. 2016. Civic complaint application under smart city project, International Journal of Advanced Computing and International Technologies, vol. 3, no. 2.

[5] Zhang, Z.; He, Q.; Gao, J. & Ni, M. A deep learning approach for detecting traffic accidents from social media data Transportation Research Part C: Emerging Technologies, 2018, 86, 580 – 596

[6] M. M. Skoric, Q. Zhu, D. Goh and N. Pang. 2015. Social media and citizen engagement: A meta-analytic review. New Media and Society, Sage Publishing.

[7] Atreja, S.; Aggarwal, P.; Mohapatra, P.; Dumrewal, A.; Basu, A. & Dasgupta, G. B. Citicafe: An Interactive Interface for Citizen Engagement 23rd International Conference on Intelligent User Interfaces, ACM, 2018, 617-628

[8] Twitter Seva https://blog.twitter.com/official/en_in/a/2016/the-ministry-of-communication-adopts-twitter-seva-in.html

[9] https://docs.aws.amazon.com/index.html#lang/en_us (accessed January 29, 2019)

[10] Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A.A review of feature selection methods on synthetic data Knowledge and Information Systems, 34 (2013), pp. 483-519

[11] Webb, G. I., J. Boughton, and Z. Wang (2005). Not So Naïve Bayes: Aggregating One-Dependence Estimators. Machine Learning 58(1). Netherlands: Springer, pages 5-24.

[12] Rojas, R. (2009). AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. Freie University, Berlin, Tech. Rep.

[13] Taggart, Allison J; DeSimone, Alec M; Shih, Janice S; Filloux, Madeleine E; Fairbrother, William G (2012). "Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo". Nature Structural & Molecular Biology. 19 (7): 719–721.

[14] R. Collobert and S. Bengio (2004). Links between Perceptrons, MLPs and SVMs. Proc. Int'l Conf. on Machine Learning (ICML).

[15] Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JB (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization". Journal of Chemical Information and Modeling. 46 (6): 2412–2422.

[16] Fradkin, Dmitriy; and Muchnik, Ilya; "Support Vector Machines for Classification" in Abello, J.; and Carmode, G. (Eds); Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, volume 70, pp. 13–20, 2006

[17] https://montrealgazette.com (accessed September 1, 2018)