EGOV-CeDEM-ePart

Sentiment in Citizen Feedback: Exploration by Supervised Learning

Robin Lybeck*, Samuel Rönnqvist**, Sampo Ruoppila***

*Sociology, Åbo Akademi University <rlybeck@abo.fi
**Turku NLP, University of Turku <saanro@utu.fi
***Social Research, University of Turku <sampo.ruoppila@utu.fi

Abstract: Web-based citizen feedback systems have become commonplace in cities around the world, resulting in vast amounts of data. Recent advances in machine learning and natural language processing enable novel and practical ways of analysing it as big data. This paper reports an explorative case study of sentiment analysis of citizen feedback (in Finnish) by means of annotation with custom categories (Positive, Neutral, Negative, Angry, Constructive and Unsafe) and predictive modelling. We analyse the results quantitatively and qualitatively, illustrate the benefits of such an approach, and discuss the use of machine learning in the context of studying citizen feedback. Custom annotation is a laborious process, but it offers task-specific adaptation and enables empirically grounded analysis. In this study, annotation was carried out at a moderate scale. The resulting model performed well in the most frequent categories, while the infrequent ones remained a challenge. Nonetheless, this kind of approach has promising features for developing automated systems of processing textual citizen feedback.

Keywords: e-participation, machine learning, big data, sentiment analysis

1. Introduction

Web-based e-participation tools have become a conventional way for citizens to voice their initiatives and concerns to city administrations in many cities around the world. These tools range from specific map-based questionnaires (Kahila-Tani 2015) to general citizen feedback services allowing citizens to submit messages to the local government (Evans-Cowley and Hollander 2010). Consequently, volunteered geographic information on the urban environment is generated and saved in vast amounts on the back-end. It is generally used by city officials only once (i.e., further, iterative analysis is conducted but rarely). One reason for this has been the unsuitability of any conventional methods of text analysis, due to the very large size (and continued rapid growth) of the data. Therefore, there has been a great interest in applying advanced computational methods with a data-driven approach (Münster et al. 2017).

In this paper, we explore the tone and sentiment of citizen feedback data, collected from a particular service by means of supervised machine learning regularly used in Turku (Finland) since 2012. We combine the viewpoints of computer science and social sciences, taking an explorative approach. We use a supervised learning method developed in the field of natural language processing, and we analyse the results qualitatively to evaluate the outcome. The corpus of the data consists of various types of citizen feedback regarding the urban environment and place-based development. In a broader research project, of which this study is a part, the sentiments and thematic information are integrated with location-based analysis.

We explore the data using a custom scheme of sentiment categories deemed appropriate in the context of citizen participation in urban planning. The expression of sentiments is highly context-dependent, meaning that sentiment analysis tools may work well only within specific domains. Therefore, creating a dataset customized to the domain and task at hand may provide more accurate and meaningful results. Custom annotation of data may also be necessary in order to process text in non-major languages (such as Finnish) that lack available linguistic resources. Using an empirical approach to modelling allows for exploration of the material, which may reveal unanticipated patterns of expression of sentiments by citizens.

Analysing the sentiment, tone and expression of emotions in texts has also been of interest in previous e-participation research (Liu 2012; DiMaggio 2015). In particular, sentiment analysis on public social media platforms has been found as a way of measuring public opinion (Thelwall 2014; Gonzalez-Bailon and Paltoglou 2015). Sentiment in social media communication has been examined in relation to citizen participation and popular attitudes toward government and policy (see, e.g., Zavattaro et al. 2015 and Arunachalam & Sarkar 2013). Focusing particularly on citizen participation in urban planning, Münster et al. (2017, 2400–2401) have highlighted the challenges of extracting relevant information through sentiment analysis and other computational text-analysis methods.

Methodologically, many previous studies rely on dictionaries of sentiment-bearing words, which have been considered insufficient in capturing the broad variety of how sentiments are commonly expressed (Liu 2012). In fact, methods based on word frequencies and dictionaries have been used in text analysis for quite some time (Gonzalez-Bailon and Paltoglou 2015; Rykov et al. 2016). Another method is to use a data-driven, empirical approach to discover unanticipated information and ways of expression, such as through analysis of patterns of word frequencies (DiMaggio 2015). However, neutral sentences that contain sentiment words or opinionated sentences that lack clear sentiment words are problematic (Liu 2012). Furthermore, while dictionaries may be a quick way to find meaningful information, their use relies on the knowledge and intuition of the person(s) constructing the resource, which limits the coverage. Another problem with both types of analysis is that they operate on the level of individual words, largely disregarding meaning in context. By treating words as distinct symbols, these methods may also have difficulty in conveying the prevalent variant in language usage. A more reliable – albeit more laborious – approach, applied in this paper, is to instead annotate the text containing expressions of sentiment, using machine learning to infer how different sentiments are expressed.

In the last few years, neural networks and deep learning have emerged as particularly promising methods for many tasks of pattern recognition and artificial intelligence, including computational text analysis (Schmidhuber 2015). Neural networks, a type of machine learning, have become widely popular in natural language processing due to their ability to flexibly and more accurately model semantically meaningful patterns in text.

The two main types of machine learning are supervised and unsupervised learning. In supervised learning, the aim is to train a model that can estimate a certain variable based on other information in the data, such as positive or negative sentiments of a comment, based on its words. This requires data where the target variable has been labelled for several instances, which can be used to train the model and evaluate its predictive performance. In order to obtain such data, it is often necessary to manually annotate the set. In text analysis, annotating a corpus oneself allows modelling of a particular type of material or aspect, and it may lead to more targeted and meaningful analysis, but it is also a laborious process. Unsupervised learning does not require labelled data but instead aims at identifying naturally occurring patterns, and it can be used to model semantics in large corpora and improve prediction when annotated examples are relatively scarce.

In the rest of the paper, we describe the data and the process of annotating a citizen feedback corpus, present a machine-learning model trained to evaluate it, demonstrate how the model performed, and conclude by discussing the methodological challenges of this approach.

2. Research Data and Methods

The research data consists of 22,000 "messages" (i.e., textual feedback written by citizens, drawn from Turku's official web-based feedback channel (https://opaskartta.turku.fi/eFeedback/)). The messages analysed were submitted between January 2012 and April 2017. During that time period, a total of circa 70,000 messages were submitted, from which we categorically selected messages focusing on the urban environment or place-based development; they also had either GPS coordinates or an address. The issues typically not related to spatial development, such as feedback on healthcare services or educational institutions, were excluded. Moreover, we did not include messages not made public by the users.

Over the years, the feedback webpage has been developed from having a relatively open-ended design to a more structured service where the user is expected to categorize the feedback and encouraged to provide a location. The design encourages reporting of specific issues rather than giving general ideas, and the messages submitted are mostly short. The data were processed to identify and delete outliers (e.g., duplicates and false categorizations), as these messages could have a negative impact on the accuracy of the model.

For the purpose of this study, we annotated part of the data manually with a system of codes used for model training and evaluation. The annotated categories consisted of sentiment/tone (Positive, Neutral, Negative, Angry and Constructive). For further analysis, we also annotated a thematic tone category of (people arguing something to be) unsafe. We chose to take an

empirically grounded, data-driven approach to the material. The data were coded in three phases. The material was first explored with some descriptive analyses (word frequencies and keywords) to obtain an overview. Furthermore, we also explored it using existing thematic structures (i.e., categorizations provided by the municipality) in order to assure its usability (in further analysis). We coded a substantial amount of training data (n=1779) in this way, guided by the codebook shown in Table 1. In the second phase, we coded one thematic category (Unsafe) automatically based on codewords (n=1452). In the third phase, we extended the set by 750 randomly sampled messages and coded them in order to improve coverage and model performance for the less frequent categories. The resulting 3981 messages used in our reported experiments were shuffled and split into a training set (70%) and a test set (30%).

Table 1. Codebook

Code	Description	Example
Positive	The message reports that something positive has occurred.	The streetlight was fixed. Thank you!
Negative	The message conveys that something is wrong and implies that someone is to blame.	Despite my earlier messages, the streetlight is still not working (implying it should have been fixed).
Neutral	The message does not contain any sentiment.	This streetlight does not work.
Angry	The message contains strong language.	The streetlight is still not working. Why don't you fix it, you lazy bastards!?!
Constructive	The message contains a suggestion of how to improve something that works or fix something that is broken.	The streetlight broke five times during the last year. The municipality should switch to LED lights, which are much more durable.
Unsafe	The argument in the message reports that something in the urban environment is unsafe or dangerous.	A lot of dangerous situations happen because of a streetlight not working. One cannot see anything!

The annotation codes for the categories were defined according to a scheme based on the implicit and explicit content of the messages (see Table 1). These categories were chosen, as they provide insight into the ways that citizens use this type of channel to communicate issues. We expanded the conventional categorization of positive, negative and neutral sentiments to include analysis of angry and constructive messages, as well as messages containing wording implying an unsafe situation. The Constructive category indicates that the person providing the feedback also wanted to contribute to solving the problem or improving a working system. This sentiment or tone implies a higher level of agency, compared to the other categories. The Angry category informs how much of the feedback contained hateful or aggressive language. The semantic category labelled Unsafe was coded for all messages containing a comment on something being an

issue because of safety concerns. The categories were coded non-exclusively (i.e., a message could belong to several categories at the same time).

We constructed a neural network as a predictive model, consisting of multiple interconnected layers where information is passing from input to output while being transformed according to trainable parameters adjusted through supervised learning. We adjusted these internal parameters in order to achieve optimal correspondence between the output of the network model and the labels of the data. The network consists of the following layers: 1) an input layer (representing input words as discrete symbols), 2) an embedding layer (providing continuous vector representation of word meaning), 3) a recurrent layer (traversing words in sequence and providing a single vector representation of a whole comment, implemented using Bidirectional Long Short-Term Memory (Graves and Schmidhuber 2005)), and 4) an output layer (transforming vector representation into probabilities over the target variable labels, using sigmoid activation). The parameters of the embedding layer were initialized through unsupervised learning, namely by-word vector representations obtained by applying the word2vec skip-gram algorithm (Mikolov et al. 2013) on the whole corpus, as well as on other unannotated texts in the same language. This provides better coverage and generalization of word meanings in the model. The rest of the network was trained on the annotated comments and provides a baseline that can demonstrate how the corpus is able to support analysis of sentiment and themes in any similar text.

3. Results

Having trained the predictive model, we evaluated it quantitatively by means of the test set to assess how well it learned to recognize the coded categories for the held-out data. This provided quality assurance based on a larger number of messages. After that, we continued by analysing the output qualitatively, in order to evaluate how the model performed.

3.1. Quantitative Evaluation Results of the Model

The results are presented in Table 2 with a breakdown per category. The *F1-score* represents a harmonic mean of *precision* (correctness rate of identifications) and *recall* (rate of actual categorizations identified), whereas *support* shows the number of positive examples of each category in the test set. The substantial variation of the frequency per the categories is largely correlated with the predictive performance, as more training examples let the model generalize better to a specific category. In particular, the model performs well for the Unsafe, Negative and Neutral categories with F1-scores above 60 percent, and the Angry category is also well recognized, although infrequent in the data, due to explicit expressions. Overall, the model can predict with an F1-score of 70 percent (88.4% accuracy).

Table 2. Predictive Performance of the Model on the Test Set

F1-score	Precision	Recall	Support	Category
17.7%	100%	9.7%	31	Positive
63.5%	61.1%	66.1%	221	Neutral
64.5%	58.4%	72%	382	Negative
40%	57.1%	30.8%	26	Angry
30.7%	38.5%	25.5%	98	Constructive
84%	76.8%	92.8%	553	Unsafe
70%	66.4%	74%	-	Overall

3.2. Qualitative Analysis of Representative Cases and Distinctive Keywords

The categorization provided by the model on unannotated data was used to expand and organize the material for qualitative analysis. By categorizing the rest of the material, we extracted keywords to describe each category based on more examples (using word-frequency-based TF*IDF weighting). We also used the model to select representative messages for each comment, based on the model's confidence in its decision (posterior probability for the selected class), presented in Table 3.

Table 3. Examples of Representative Cases Retrieved by the Model (Authors' Translation)

Sentiment	Retrieved comment	Confidence
Positive	"It was nice to note that in Huhkola the walking path between the Hiihtomajapolku road and Näädänkatu was sanded and the area was cleaned. Thank you!"	0.725
Neutral	"Next to the jogging track (on the right side when coming from the underpass), a big pine tree is leaning somewhat over the jogging track. It is quite close to the speed bump sign on Aurakatu."	0.929
Negative	"In Marinkatu, Vaala, no sand has been removed from the sidewalk or street despite the city's maintenance decision, of which the residents have been informed in a notice dated 13 May 2011. In addition, in the whole area – Marinkatu, Ugrinkatu, Mordvankatu and Vatjankatu – no sand has been removed. The snow has melted a long time ago and from nearly everywhere else it has already been removed! Why is this residential area always the last one to get the maintenance service of the streets?"	0.999
Angry	"Outrageous! Bus no. 30, at the intersection of Merimiehenkatu and Stålarminkatu (Stop 101), to the city, at 9:34 am!! The driver did not take my daughter, who was going to school, nor anyone else at the stop waving at the bus!! She went to school in the rain and was late! CELL PHONES SHOULD BE REMOVED FROM THE DRIVERS! That driver was talking on a cell phone. IT IS NOT THE FIRST TIME THAT THE CUSTOMERS ARE LEFT TO WAIT FOR THE NEXT BUS OR THE BUS DOES NOT STOP TO LET YOU OUT!"	0.864

Sentiment	Retrieved comment	Confidence
Constructive	"Would it be possible to get some kind of lights for Barker's playground? The park is very popular in the evenings and the streetlights on the promenade do not illuminate the play area adequately. The fastest solution would be to attach the lights to the streetlight poles at the end of the park by Rantakatu and turn them to face the playground. Thank you for the good playgrounds in the city centre!"	0.883
Unsafe	"By Myllylahti, the zebra crossing between the overhead bridge and the tunnel causes many dangerous incidents. A three-lane road, lots of traffic and a small schoolboy on the edge of the road wanting to cross the road. Impossible for drivers to have visibility of the pedestrian from the lane furthest away. Is it necessary to have a zebra crossing in this place when there is an overpass for pedestrians before the tunnel, and after a tunnel there are crossroads which naturally have a zebra crossing? However, if this zebra crossing is necessary, it would be an absolute priority, especially for improving road safety for schoolchildren, to install flashing lights at the zebra crossing, as elsewhere.	0.999

We also identified the most distinctive keywords in each category, based on the model's prediction. The keywords were used to support a qualitative reading of 10 comments of each category to confirm that the model works reasonably well, at least in the case of comments with a relatively high level of confidence (0.72 or higher).

The Positive category performed relatively well, based on the qualitative reading. All comments selected by the model (except one) contained a positive sentiment. The wording in that one comment was positive, but the meaning was sarcastic. The most distinctive words confirmed that the category works relatively well, as ten out of fifteen were clearly positive. Interestingly, the Neutral category had the weakest results when analysing the output qualitatively, although it performed quite well quantitatively. This was confirmed by the most distinctive keywords, where, for instance, street names appeared. This type of occurrence indicates that either the training data had too few examples annotated in this category or the annotation was not done very consistently. The choice to work within the structure of the data in the first phase of the annotation process may have resulted in skewed annotation data. However, the problem seems to have been limited to this one category. The Negative category performed very well, with all the comments containing clearly negative sentiments. This was confirmed by the most distinctive keywords, where the most frequent word was translated as "not taken care of" (230 cases). In the Angry category, all comments included some form of strong language, and most of them also contained other ways of expressing strong sentiment, such as use of the caps lock or exclamation marks. The distinctive keywords were almost exclusively comprised of swearing. The Constructive category also performed well, with the most distinctive keyword being translated as "would it be possible". Finally, the Unsafe category, which was the most frequently coded in the material, performed flawlessly from a qualitative perspective. The overall conclusion of the qualitative reading is that although the model performs relatively well, annotation of an even larger set of training data

would be required to obtain satisfactory performance, especially in less frequent categories. The laborious process of annotating the corpora is justified, as the approach could pay off in the long run for the purposes of developing automated monitoring systems or conducting large-scale research.

4. Discussion and Conclusion

In e-participation, a crucial question is how to enable the ideas, thoughts and experiences of citizens to be aggregated into a format that can be easily taken into consideration in the decision-making processes. The methods from computer science and natural language processing have a lot to contribute, both in terms of research and for developing the use of participatory data in local governance. Thanks to the implementation of various easy-to-use e-participation tools, the amount of citizen feedback received by municipalities has been growing exponentially. Furthermore, citizen feedback is also provided on social media and microblogging platforms, apart from formal participation channels.

Currently, city officials tend to process all citizen feedback on a one-off basis. However, by approaching it also as big data, planners would benefit from a plenitude of analytic perspectives making use of sentiments, thematic categories, spatial locations and time periods. Indeed, the smart city paradigm, involving the vision of real-time aggregated data analytics, should also include data collected through citizen participation.

Automated text analysis can serve as a useful tool in handling this mass of information. To obtain this goal, high reliability of the model's accuracy is required, combined with channelling unclear cases to a human moderator. Explorative studies, such as this one, can have both analytical relevance for method development (Bone et al. 2016) and practical benefits for improving processing of citizen feedback.

The performance of a machine-learning algorithm is directly connected to the extent and quality of the training data, as any skew will be replicated by the model. The results can be improved by annotating more material and running qualitative follow-up studies, like those presented in this paper. E-participation research has much to gain from this type of method, yet significant resources are required to construct sufficiently well-performing models. Moreover, alternative machine-learning approaches based on representations in vector space, rather than dictionaries, can achieve more nuanced output. In that case, too, the robustness of the training data and the scope of the annotations required need to be substantial in order for the model to also perform well with less frequent categories.

When employing the model in practice, the risk of misclassification should be accounted for, as well as attention to minimizing either false positives or false negatives. For example, sarcasm is a type of sentiment that is particularly tricky to identify, and it illustrates the difficulty of processing human language with great accuracy. Considering these issues, machine-learning methods best serve as a means of assisting in data analysis, rather than replacing the need for qualitative reading by officials themselves.

Our study has shown that an empirically grounded approach has the potential to achieve a good level of accuracy, as it is able to capture varying patterns of actual language use. Further annotation of the material is only expected to increase the performance quality of the less frequent categories, but it is still very laborious. Nonetheless, the predictive model can provide a starting point for exploratory and qualitative analysis of big data sets. Machine-learning methods have enormous potential in this regard, although further research is needed to improve methodological frameworks and practices before machine learning can become a standard part of research designs. Nevertheless, this type of model has promising features for developing automated systems of processing textual citizen feedback.

References

Arunachalam, R., & Sarkar, S. (2013). The New Eye of Government: Citizen Sentiment Analysis in Social Media. IJCNLP 2013 (SocialNLP), pages 23–28, Nagoya, Japan, 14 October 2013.

Bone, J., & Emele, CD, Abdul A, et al. (2016). The social sciences and the web: From "Lurking" to interdisciplinary "Big Data" research. Methodol Innov 9:205979911663066. doi: 10.1177/2059799116630665

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). Big Data Soc 2:205395171560290. doi: 10.1177/2053951715602908

Evans-Cowley, J., & Hollander, J. (2010). The New Generation of Public Participation: Internet-based Participation Tools. Plan Pract Res 25:397–408. doi: 10.1080/02697459.2010.503432

Gonzalez-Bailon, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. Ann Am Acad Pol Soc Sci 659:95–107. doi: 10.1177/0002716215569192

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. Proc Int Jt Conf Neural Networks 4:2047–2052. doi: 10.1109/IJCNN.2005.1556215

Kahila-Tani, M. (2015). Reshaping the planning process using local experiences: Utilising PPGIS in participatory urban planning. Aalto University.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. 1–108. doi: 10.2200/S00416ED1V01Y201204HLT016

Mikolov, T., & Chen, K., & Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. 1–12. doi: 10.1162/153244303322533223

Münster, S., & Georgi, C., & Heijne, K., et al. (2017). How to involve inhabitants in urban design planning by using digital tools? An overview on a state of the art, key challenges and promising approaches. Procedia Comput Sci 112:2391–2405. doi: 10.1016/j.procs.2017.08.102

Rykov, Y., & Kremenets, A., & Cerrone, D., et al. (2016). Semantic and Geospatial Mapping of Instagram Images in Saint-Petersburg City.

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks 61:85–117. doi: 10.1016/j.neunet.2014.09.003

Thelwall, M. (2014). Sentiment Analysis and Time Series with Twitter. In: Twitter Society.

About the Authors

Robin Lybeck

Robin Lybeck is a doctoral candidate at the sociology department in Åbo Akademi University. He has previously worked on issues relating to mobile participation and is writing a PhD about e-participation in urban planning.

Samuel Rönnqvist

Samuel Rönnqvist is a postdoctoral researcher in the TurkuNLP group at the University of Turku, and a visiting researcher in the Applied Computational Linguistics Lab at Goethe University Frankfurt. He holds a PhD in Computer Science from Åbo Akademi University. His research centers around language technology and deep learning, as well as their interdisciplinary application.

Sampo Ruoppila

Dr Sampo Ruoppila is Research Director of urban studies at the University of Turku, Finland, and director of Turku Urban Research Programme, which promotes research-based policy advice for the local municipality. His research interests include urban policy and e-participation in urban planning.