Amsterdam
University
Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.1 (2023) 1–26
HTTPS://DOI.ORG/10.5117/CCR2023.1.10.LIAN

Word embedding enrichment for dictionary construction: An example of incivility in Cantonese

Hai Liang
School of Journalism and Communication, The Chinese University of Hong Kong, HKSAR

Yee Man Margaret Ng
Department of Journalism, University of Illinois Urbana-Champaign, USA

Nathan L.T. Tsang
Department of Sociology, The University of Southern California, USA

Abstract
Dictionary-based methods remain valuable to measure concepts based on texts, though supervised machine learning has been widely used in much recent communication research. The present study proposes a semi-automatic and easily implemented method to build and enrich dictionaries based on word embeddings. As an example, we create a dictionary of political incivility that contains vulgarity and name-calling words in Cantonese. The study shows that dictionary-based classification outperforms supervised machine learning methods, including deep neural network models. Furthermore, a small number of random seed words can generate a highly accurate dictionary. However, the uncivil content detected is only weakly correlated with uncivil perceptions, as we demonstrate in a population-based survey experiment. The strengths and limitations of dictionary-based methods are discussed.

Keywords: political incivility, machine learning, dictionary construction, Cantonese, swearing

Introduction

The availability of large-scale social media textual data has helped computer-assisted content analysis flourish. Given the large volume of the datasets, manual coding in conventional content analysis has become less popular and is not even feasible in some contexts. Researchers have turned to computational tools for large-scale text mining to supplement or replace traditional content analysis to measure social constructs based on texts

(Grimmer et al., 2021). Although supervised machine learning methods have been widely used in measuring communication concepts like sentiment, incivility, and topic of text content (e.g., Theocharis et al., 2016; van Atteveldt et al., 2021; Wojcieszak et al., 2023), dictionary-based methods remain valuable in some contexts (e.g., Dun et al., 2021; Guo et al., 2016; Muddiman et al., 2019). However, building dictionaries can be time-consuming and expensive, and the procedures are usually less systematic than supervised machine learning. The present study proposes a semi-automatic and easily implemented method to build a political incivility dictionary that includes vulgarity and name-calling words based on word embeddings. This study demonstrates that, in this context, the dictionary-based classification approach outperforms supervised machine learning.

The study contributes to the literature in several ways. First, the present study proposes a convenient method to generate keywords not only for measurement but also for information retrieval. Previous studies relied on domain experts or crowdsourced coders to construct dictionaries; however, the recruitments and training are usually expensive and time-consuming (Fast et al., 2016; Mohammad & Turney, 2010). Crowdsourced coding could also be problematic when the domain involves very informal languages like our corpus from web forums in Cantonese. Several systematic and automated methods have been proposed to cope with these problems. For example, researchers have built dictionaries by performing label propagation over word co-occurrence or similarity graphs (e.g., Hamilton et al., 2016; Velikovich et al., 2010). Following this tradition, the present study proposes a systematic procedure to generalize this approach to dictionary construction. Instead of using label propagation algorithms, the proposed procedure checks relevant words manually and iteratively. While human validation is still necessary, validating individual words is less time-consuming than coding entire sentences, paragraphs, or articles as required in supervised machine learning. The study also tested the efficiency and accuracy of selecting seed words and found that just a few iterations with 10 randomly selected words can achieve high accuracy.

Furthermore, the keywords generated through this method can also be used for information retrieval purposes, such as searching for relevant documents on specific topics from a large text corpus. In some cases, it may not be feasible to obtain a random sample of documents to create a training dataset for machine learning (e.g., a random sample of Facebook comments), or even if a random sample is obtained, it may require a very large sample to ensure sufficient target cases (e.g., hate speech in an election). In such situations, a comprehensive list of keywords can be used to retrieve relevant documents for further studies. While similar methods have been proposed for this purpose in other research contexts (e.g., King et al., 2017; Tong et al., 2022), the proposed method here is easily implemented and does not require a large human-coded training dataset.

Second, the present study demonstrates that complicated models are not always better than simple ones. Although dictionary-based methods are relatively convenient and cost-efficient, their performance is highly context-specific (González-Bailón & Paltoglou, 2015). Theoretically speaking, when supervised machine learning is trained on a large enough random sample, it would outperform dictionary-based methods (Barberá et al., 2021). Empirical comparisons support the view that supervised machine learning is usually better than dictionary-based methods in terms of accuracy (e.g., van Atteveldt et al., 2021; Widmann & Wich, 2022). Nevertheless, the present study shows that, in practice, dictionary-based methods could outperform machine learning and even crowdsourced coding.

Third, most existing studies analyze English and other Western language corpora (Baden et al., 2022), while the present study focuses on the less frequently studied language of Cantonese, which is among the most influential dialects in the Sinitic (Chinese) languages and is spoken by about 85 million people, most of whom are in southern China, Hong Kong, and Macau. Although Chinese lexicons in Mandarin are available, lexical Cantonese databases are relatively rare. Those Cantonese dictionaries that do exist, such as the Hong Kong Cantonese Corpus (Luke & Wong, 2015) and Cifu (Lai & Winterstein, 2020), were designed for specific domains. These corpora are not online sources and do not claim to cover user-generated content on social media. Therefore, the proposed method is not just useful for general dictionary construction; it also serves as one of the few large-scale empirical studies of Hong Kong's online discourse.

Literature review

Dictionaries vs. supervised machine learning

The dictionary-based approach is likely the most commonly used automated content method in the social sciences due to its efficiency, transparency, and simplicity. This approach uses a dictionary—a collection of pre-sorted words—to define categories. Textual data are compared to dictionaries using a variety of metrics (e.g., the number of times the words appear in each document), and these metrics assign the text to specific categories.

Some off-the-shelf dictionaries, such as Linguistic Inquiry and Word Count (Pennebaker et al., 2007, LIWC:), Lexicoder Sentiment Dictionary (Young & Soroka, 2012, LSD:), and the Moral Foundation Dictionary (Graham et al., 2009) have been widely used to classify textual sentiments, recognize theory-informed constructs, and extract moral intuitions in media framing.

However, a well-performed dictionary analysis relies heavily on the existence of suitable dictionaries in specific domains and languages (Grimmer & Stewart, 2013), which may fail to consider community-specific vernacular or demographic variations in language use (Hovy, 2015; Yang & Eisenstein, 2015). In large-scale text analysis, researchers create a list of keywords to retrieve target documents from the population (King et al., 2017). For example, researchers usually select a few hashtags to retrieve tweets related to political elections or social movements via Twitter's streaming API. Thus, the accuracy and validity of the results are highly dependent on keyword selection. The agreement was generally close to a chance agreement, and correlations between dictionaries were also low (e.g., Young & Soroka, 2012). Error analysis showed that this was mostly due to the missing context of words (van Atteveldt et al., 2021).

Previous studies have suggested several methods to validate content dictionaries in specific contexts (e.g., Chan et al., 2021; Muddiman et al., 2019; van Atteveldt et al., 2021). If an established dictionary fits a specific task, it is straightforward to apply it after appropriate validation. However, if an appropriate dictionary does not exist, researchers have to create one from scratch, which might be time-consuming. Therefore, it is important to come up with a systematic procedure to build a context-dependent dictionary in a fast and convenient way. To extend previous research, the present study proposes a method that could generate keywords semi-automatically and validate them manually at the same time.

In contrast, supervised machine learning methods do not depend on a list of pre-sorted words but on word context and patterns. The method requires researchers to prepare a sufficiently large amount of annotated data—documents that are either labeled manually by human coders or automatically by, for example, dictionaries. Raw documents are transformed into features such as word counts, term frequency-inverse document frequency (TF-IDF) scores, and topic probability scores derived from topic modeling, which can be used as input for the analysis. A portion of the annotated documents will become training data that are used to create an algorithm to learn the relationships between the selected features and the annotations. Finally, researchers validate the model accuracy with test data, another portion of the annotated documents, with correct answers. Once a certain performance level in terms of precision, recall, and F1-scores is reached, the classifier can be used to annotate other unseen documents. Although machine learning models might identify spurious patterns in the data and are also content-specific (Thelwall et al., 2010), they have been demonstrated to generally be more accurate than off-the-shelf dictionaries in measuring emotional language in political and economic discourse (van Atteveldt et al., 2021; Widmann & Wich, 2022).

However, these two approaches to automated content analysis need not compete; they can complement each other, and previous scholarship suggests that supervised learning augments can extend dictionaries and vice versa. Dun et al. (2021) applied hierarchical dictionary counts and supervised learning (trained on sentences extracted using dictionaries and coded manually) jointly to measure media coverage of changes in U.S. defense spending. They found that the combined approach performed slightly better than the dictionary alone. Following Dobbrick et al. (2022), Jakob et al. (2023) combined LIWC dictionaries with machine learning to measure toxic outrage in user comments on Facebook, Twitter, and news website comment sections. These studies advocated producing a training data set using a dictionary-plus-supervised-learning approach since pre-processing the data with off-the-shelf dictionaries reduces the quantity of annotated data that would be needed for full-text machine learning.

Measuring political incivility

Specifically, we take political incivility as an example to present our dictionary enrichment method. Although there is a lack of uniformity in the literature on the definition of incivility, the frequently used definition by (Coe et al., 2014, p.660), which we adopt for this study, describes incivility as "features of discussion that convey an unnecessarily disrespectful tone toward the discussion participants or its topics." Vulgarity is perhaps the most recognizable form of incivility. It is usually exhibited as profanity or foul language, curse words, or certain taboo words that are "generally considered inappropriate in professional discourse." Name-calling is another damaging form of insults or attacks, which are "directed at a person or group of people" (Coe et al., 2014, p.660), usually by labeling the targeted individual or group with a pejorative and demeaning name. According to this definition, bad labels directed at non-human objects or entities are not considered name-calling. Therefore, disparaging words directed at organizations or companies were not considered uncivil in the present study.

Past scholarship has used both supervised machine learning and dictionary-based approaches to detect political incivility in news and social media texts. Supervised machine learning approaches recognize the unique features of incivility from labeled data and ultimately learn to identify these features in unlabeled comments. For example, using supervised machine learning, Theocharis et al. (2016) detected uncivil comments on Twitter. They manually coded a random sample of tweets along two dimensions: politeness versus impoliteness and whether a tweet contained a reference to moral or democratic issues. They selected regularized logistic regression to train the models for the two dimensions separately, eventually achieving an overall accuracy above 0.80. If a tweet was impolite and related to morality or democracy, it was classified as uncivil.

Stoll et al. (2020)'s study on impolite and uncivil comments on German media's Facebook pages found that traditional classifiers (naïve Bayes, decision trees, support vector machines, and logistic regression) could only measure comments at the word level; the best model performances were achieved by naïve Bayes (accuracy = 0.64). This is mainly because these classifiers cannot detect subtle forms of incivility or predictive words of incivility or impoliteness that were used in non-offensive ways. Timm and Barberá (2019) studied incivility on U.S. legislators' Facebook pages and found that using the dictionary can flag potentially harmful comments that supervised models do not pick up, possibly due to small-scale training materials.

Dictionary-based methods perform well only when the concept analyzed is closely related to the word level of a statement. On the one hand, predefined word lists cannot comprehensively measure the concepts related to incivility, missing comments that are subtly abusive without the use of profane language, such as covert racism or sexism (e.g., Cho & Kwon, 2015; Muddiman & Stroud, 2017; Stoll et al., 2020). On the other, these methods can misclassify linguistic nuances (Burnap & Williams, 2015). Similarly, T. Davidson et al. (2017) found that only 5% of the tweets containing words in the English hate speech lexicon hatebase.org were labeled as hate speech by human coders.

To refine the construction of an incivility dictionary, Muddiman et al. (2019) suggest a deductive approach to building context-specific incivility dictionaries. Starting from a top-features list, they went through several iterations of human coder validation to ensure that the texts did indeed use the features in a way that aligned with the purpose of the dictionary. The performance of this manually validated organic dictionary approach, which had an overall accuracy of 73.1%, compared favorably to human coders, other sentiment dictionaries like LIWC, and machine learning algorithms. The present study extends this method by automatically generating the top features based on word embeddings. As we explain in the next section, word embeddings can help find synonyms and analogies and thus improve the recall of a dictionary. It also increases the replicability of the dictionary-building process.

All automated methods, however, decline in performance when applied to a different semantic task or domain (Muddiman et al., 2019; van Atteveldt et al., 2021). This makes it difficult to estimate beforehand which method offers the best accuracy and is most cost-effective. Moreover, despite the wide availability of advanced computational techniques, those resources are disproportionately focused on European languages, especially English, sidelining other languages spoken globally. As such, communication scholars often face a lack of robust language resources to conduct topic and sentiment analysis for research involving multiple languages. Researchers often need to recruit coders who specialize in different languages to implement multilingual analysis. However, this is challenging due to the cost of recruiting and training multilingual coders (Reber, 2019). Many times, scholars resolve this challenge by translating their original corpus into English using Google Translate API (de Vries et al., 2018; van Atteveldt et al., 2021), which undoubtedly raises questions about the semantic correctness of the translations and the precise preservation of the meanings associated with the original text.

Word embeddings for text classification

Conventional supervised machine learning and dictionary-based models are usually based on bag-of-words features. The recently developed word embeddings and deep learning models have been incorporated to improve the performance of both supervised and unsupervised models. One of the most basic text-based features is the bag of words, which takes the occurrence of words as input features. However, this approach ignores aspects of word order and grammar. More recently, word embeddings, which consider the relationships between words and the communication context, have been developed to address this limitation. Neural network models are generally used with word embedding models in which words are represented as vectors and can be described as the relative location of a word in an n-dimensional vector space (Goldberg, 2017). To arrange the words in the vector space, words from an extensive corpus of documents are fed into a neural network and mapped (embedded) into lower-dimensional vector representations. In this representation, words used in similar contexts have similar vectors. This approach has proven superior to text classification models that process each word separately using the bag-of-words model (Devlin et al., 2019). In particular, deep neural network classifiers like convolutional neural networks generally outperform dictionaries in state-of-the-art systems (Rudkowsky et al., 2018; van Atteveldt et al., 2021). This is because of the higher learning capacity possessed by deep neural networks with multiple hidden layers. However, the complexity inherent in many deep-learning approaches usually poses problems with interpretability and transparency.

Word embeddings must be trained separately on enormous numbers of text documents. Therefore, researchers often use pre-trained word embeddings, which should be trained on a dataset that is comparable to the dataset that will be classified. Widmann and Wich (2022) demonstrate that transformer-based models that come with pre-trained language models outperform off-the-shelf dictionaries and simple neural network classifiers in classifying emotional language in German political discourse. In addition, the authors present a method to augment existing sentiment dictionaries based on word embeddings; their approach outperforms the original dictionaries. However, few pre-trained models are available for classification tasks using non-English and non-formal language texts such as online discussions. The present study trains a word embedding model based on a massive number of online comments in Cantonese.

Word embeddings and deep learnings have been applied to classify incivility-related content. For example, Rudkowsky et al. (2018) present a process pipeline of supervised sentiment analysis with word embeddings to estimate levels of negativity in Austrian parliamentary speeches; they achieved an average accuracy of 0.58. S. Davidson et al. (2020) compared how Bidirectional Encoder Representations from Transformers (Devlin et al., 2019, BERT) and the DistilBERT-based neural model, along with a simple logistic regression model (with TF-IDF features), performed on an incivility classification task. The results show that the DistilBERT model achieves the highest F1-score of the three models, but they are all competitively accurate (all F1-scores > .78). However, annotating a dataset large enough to train a high-accuracy neural classifier from scratch is a costly and time-consuming undertaking. To provide a sufficient amount of training data, previous work on incivility has employed various data augmentation techniques, such as back-translation (Ibrahim et al., 2020) and data transformation techniques (Rizos et al., 2019).

Data and method

Political discussion corpus

We collected political comments from two major online discussion forums in Hong Kong: lihkg.com (LIHKG) and discuss.com.hk (DISCUSS). These two discussion forums play an important role in Hong Kong's public sphere. DISCUSS was established in 2003 and was once the most visited local forum in Hong Kong. It held that position until the HKGolden forum, which became LIHKG in 2016, gained popularity and replaced DISCUSS as the most popular online forum during the 2019 Anti-Extradition Law Amendment Bill Movement (the Anti-ELAB Movement or Movement below). In terms of political stance, DISCUSS had more pro-establishment users and was more diverse. In terms of education level, LIHKG required an internet service provider or college or university email address for registration. In this sense, LIHKG users tended to be more educated than DISCUSS users.

The present study focuses on political incivility in contrast to impoliteness in online conversations, as incivility is a concept related to discussions of public issues (Coe et al., 2014). Only comments posted on sub-forums related to public issues were collected via web scraping. We considered all LIHKG comments to be about public issues, given that most comments were related to the 2019 protests. For DISCUSS, we manually selected 11 sub-forums that were closely related to public issue discussion (e.g., Hong Kong and World News, Breaking News). The final dataset included 65,513,807 comments posted between June 2019 and December 2020 on LIHKG (1.5 years) and 39,745,007 comments posted between January 2011 and December 2020 on DISCUSS (10 years).

Word2Vec enrichment

Figure 1 illustrates the pipeline of the dictionary enrichment method using word embeddings. We combined all comments from LIHKG and DISCUSS into a single corpus and trained a Word2Vec model. The purpose of training the Word2Vec model was to systematically find semantically similar words related to the concepts (i.e., vulgarity and name-calling) or analogies. Word2Vec proposes two kinds of models: (1) the continuous bag of words (CBOW), which learns the representations by predicting the target word according to contextual words in all comments; and (2) Skip-gram, which predicts each context word based on the target word. We utilized the Word2Vec function from the GENSIM library in Python to fit our models. As Word2Vec training is an unsupervised task, there is no universally accepted way to evaluate the results. Our objective is to detect uncivil words using a specific set of seed words. We experimented with both CBOW and Skip-gram models. For each model, we checked the most similar words of some popular uncivil words in Cantonese to see if most of them were also uncivil. We finally chose the Word2Vec model with 250 dimensions, window = 5, using CBOW.

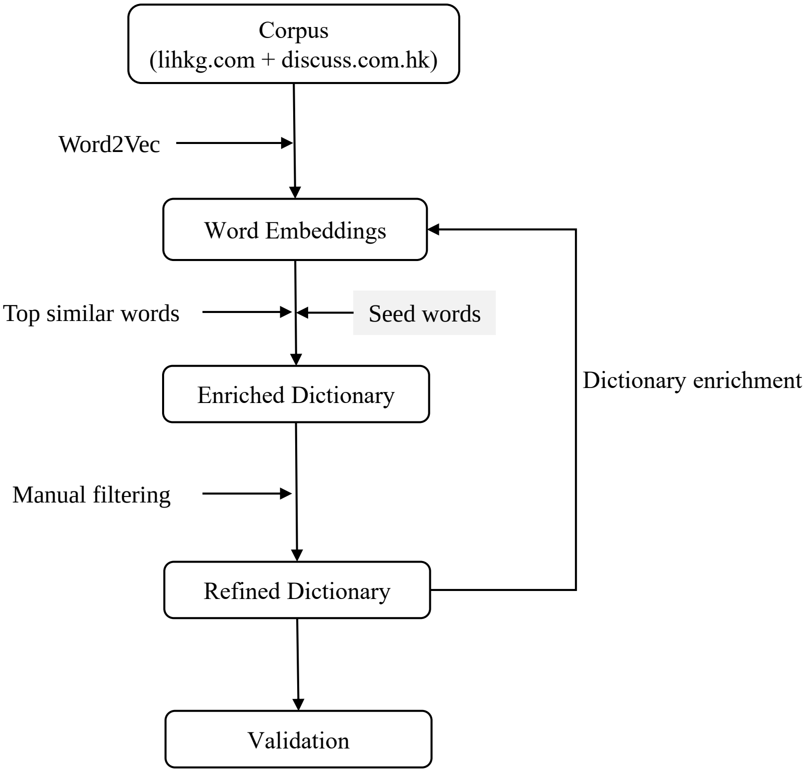

Figure 1: The pipeline of dictionary construction using word embeddings. The inputs of the method are a corpus to train word embeddings and a list of seed words related to the concepts.

Next, we compiled a list of seed words based on our knowledge of the local language. The list ($n = 307$) comprised 59 vulgar words (e.g., 撚, 屌, 柒) and 248 name-calling words (e.g., 黃屍, 支畜). At this initial stage, we collected as many unambiguous words as possible. Next, for each seed word, we used cosine similarity to retrieve the 20 most similar words based on the trained Word2Vec representations. After creating an enriched dictionary, we manually verified whether each word was related to vulgarity or name-calling. Subsequently, we used the contents of that refined dictionary as new seed words. The iterative process of dictionary enrichment was repeated until no new related words were found. As Figure 2A shows, we repeated the enrichment process 11 times for vulgarity and 15 times for name-calling. At this point, the dictionary consisted of 2,665 uncivil words (992 vulgarities and 1,673 name-calling words). Words (e.g., 碌柒) that are combinations of other words (e.g., 柒) were removed from the list. The final dictionary contains 1,956 words that could be directly applied to detect incivility in texts. The dictionary, Word2Vec, and replication materials for tables and figures are available on GitHub: https://github.com/rainfireliang/WORD-EMBEDDING-FOR-DICTIONARY-CONSTRUCTION.

Selecting seed words

The primary purpose of this study is to develop a Cantonese incivility dictionary. We started with a large number of predefined seed words ($n = 307$), which naturally raises the question of the extent to which the choice of seed words influenced the final dictionary. In practice, if the final dictionary is insensitive to the choice of seed words, a few uncivil words would be sufficient to generate a satisfactory dictionary. To answer this question, we repeated the enrichment process (as described in Figure 2A) based on n seed words ($n = 10, 50, or 100$) randomly selected from the final dictionary ($N = 1,957$). For each n, we repeated the sampling 10 times. Figure 2B presents the average number of cumulative uncivil words obtained for each round of enrichment. The results show that an n with a larger number of seed words reached its maximum sooner than an n that started with a small number of seed words. For example, 100 seed words reached the maximum in about round 6, while 10 seed words reached the maximum in about round 12, implying that the dictionary enrichment process was more efficient when starting with more seed words. In terms of the cumulative maximum, there were only minor quantitative differences after round 12. This demonstrated that even a small number of seed words could be as effective as a large number of seed words as long as the enrichment process was repeated enough times.

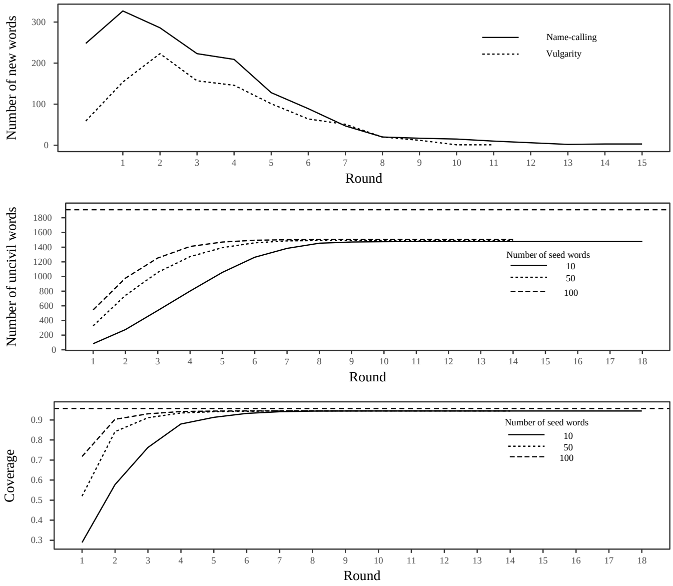

Figure 2: A (upper): the number of new words in each round of dictionary enrichment for name-calling and vulgarity, respectively; B (middle): the cumulative number of uncivil words obtained by selecting random seeds; C (lower): the coverage of uncivil incidences by selecting random seeds.

Finally, the empirical maximum was around 1,400, which was smaller than the theoretical maximum of 1,911—given that some words in our dictionary were recoded for simplicity (e.g., from 碌柒to 柒) and that we included vulgar words that were composed of alphanumeric characters (e.g., on9, 5毛) in the seed list, this meant that the theoretical maximum was 1,956 before removing 45 uncivil words from the Word2Vec representations. This missing-word problem could be resolved by using a better word embedding model like fastText (Joulin et al., 2016), which includes subwords.

On average, more than 400 words were not found in the dictionaries generated by random seeds. The reason was that the network of uncivil words was not strongly connected. Considering the topmost 20 similar relationships as network edges, we described a network as a strongly connected network if there was an edge between every two nodes (i.e., uncivil words). In the present study, the similarity network among the uncivil words had 454 strongly connected components, with the largest component featuring 1,061 nodes. This suggested that the proposed method could not completely replace experts' knowledge, which was crucial for constructing comprehensive lists of seed words.

Nevertheless, these missing words might be rare in the corpus and thus play a less important role in the detection of incivility. To justify this argument, we calculated the frequencies of all uncivil words in our dictionary and obtained the proportions by dividing the total number of words in the corpus. Then, for each round of enrichment, we calculated the proportion of the cumulative words obtained, which indicated the coverage of all uncivil incidences in the corpus. As presented in Figure 2C, the empirical maximum in terms of coverage (94.5%) was very close to the theoretical maximum (95.7%). The results indicated that the 45 words that were not in the Word2Vec representation accounted for 4.3% (100% minus 95.7%) of all uncivil incidences in the corpus, and those 400 words that were missing from the random-seed efforts account for just 1.2% (95.7% minus 94.5%) of all uncivil incidences.

Evaluation

To validate the enrichment method, two native Cantonese speakers manually coded 3,000 comments. To ensure the right balance of uncivil and civil cases, we randomly selected 1,000 online comments with vulgar words, 1,000 comments with name-calling words, and 1,000 comments without any uncivil words in the dictionary we created from all comments collected from the two discussion forums. Following the definition by Coe et al. (2014), a comment was coded 1 if it contained vulgarity or name-calling, or both; otherwise, it was coded 0. Thus, this was a binary classification task. The two coders initially coded 200 comments independently, discussed inconsistent items, and reached a consensus. After that, they coded another 100 comments to test inter-coder reliability: Cohen's kappa was 0.88, and the agreement between the two coders was 94%. The rest of the comments were then independently coded by the two coders. This manually coded dataset serves as the ground truth to evaluate the accuracy of the models; the results are reported below.

Model comparison

For comparison, we tested five supervised machine learning models that have been widely used for text classification: naïve Bayes (NB), logistic regression (LR), support vector machine (SVM), random forest (RF), and extreme gradient boosting classifier (XGB). The comments were first tokenized by *jieba*, a commonly used Chinese tokenization package, and customized stop words in Cantonese (see replication materials) were removed. After removing comments with zero tokens, our final dataset comprised 2,947 comments, with 1,707 being classified as uncivil by the two coders. The average comment length is 21 tokens ($Mdn = 10, SD = 54$). All comments were then represented as a document-term matrix and a document-term matrix with TF-IDF weighting. Terms (words) that appeared in fewer than two documents were also removed. Finally, we tested 10 classifiers with the five supervised machine learning models and the option to include the TF-IDF weighting. To achieve reliable accuracy metrics, we ran 10-fold cross-validation for each model. In each round, we randomly selected about one-third of the comments as test data (n = 973). Table 1 reports the average accuracy metrics of the 10 iterations on the test data. Using the dictionary-based method, any comment containing any words from the dictionary was labeled as uncivil. We then calculated accuracy metrics based on human annotations. Given the dictionary was not learned from a training dataset, we did not perform cross-validation.

Classifiers	Accuracy	F1-score	Precision	Recall
NB	0.70	0.76	0.84	0.70
NB+TFIDF	0.70	0.76	0.85	0.70
LR	0.76	0.78	0.75	0.82
LR+TFIDF	0.74	0.78	0.78	0.77
SVM	0.77	0.79	0.74	0.84
SVM+TFIDF	0.74	0.77	0.75	0.79
RF	0.77	0.79	0.74	0.85
RF+TFIDF	0.77	0.78	0.72	0.87
XGB	0.76	0.77	0.67	0.89
XGB+TFIDF	0.75	0.76	0.67	0.86
Dictionary	0.92	0.93	0.94	0.93

Table 1: Accuracy metrics of the classifiers used to identify political incivility.

Note. NB: naïve Bayes; LR: logistic regression; SVM: support-vector machines; RF: random forest; XGB: extreme gradient boosting classifier; TF-IDF: term frequency inverse document frequency. The number of comments used to calculate the metrics is 973.

Table 1 shows that simple supervised machine learning models can achieve reasonable accuracy, as overall accuracy ranged from 0.70 to 0.77, with RF having the highest score of 0.77 (F1-score = 0.79). Nevertheless, those models were not comparable to the dictionary-based classifier, at least in our dataset, which had the best scores in all accuracy metrics (accuracy = 0.92, F1-score = 0.93). Its overall accuracy was better than many reported in the literature, and the precision and recall scores were balanced.

One explanation of the dictionary-based method's high accuracy could be its word embeddings, which contained information from all comments on the discussion platforms, whereas the supervised machine learning models were only fed with the limited labeled training datasets. To check this argument, we further ran the five machine learning models in Table 1 by representing documents using the average of the Word2Vec scores (i.e., the average of the Word2Vec scores of all words in a document). However, the accuracy scores showed no significant improvement (see Table 2). We also ran three deep learning models—convolutional neural network (CNN), long short-term memory recurrent neural network (LSTM), and recurrent convolutional neural network (RCNN)—with the pre-trained Word2Vec as the weights in an embedding layer. RCNN had the best accuracy metrics (0.78, F1-score = 0.82) but was still not as good as the dictionary method.

Classifiers	Accuracy	F1-score	Precision	Recall
NB	0.68	0.71	0.69	0.74
LR	0.72	0.76	0.77	0.75
SVM	0.71	0.76	0.78	0.73
RF	0.79	0.82	0.83	0.81
XGB	0.78	0.81	0.84	0.79
CNN	0.73	0.78	0.83	0.74
LSTM	0.78	0.81	0.84	0.79
RCNN	0.78	0.82	0.89	0.77
Dictionary	0.92	0.93	0.94	0.93

Table 2: Accuracy metrics of the Word2Vec classifiers used to identify political incivility.

Note. NB: naïve Bayes; LR: logistic regression; SVM: support-vector machines; RF: random forest; XGB: extreme gradient boosting classifier; CNN: convolutional neural network; LSTM: long short-term memory recurrent neural network; RCNN: recurrent convolutional neural network. The number of comments used to calculate the metrics is 973.

External Validity

One of the main purposes of this study is to measure the communication concept of political incivility. Even though the accuracy of the dictionary-based method outperformed machine learning models, this only indicates that the dictionary could detect uncivil comments by their content features. However, does dictionary-based incivility reflect social reality and real-world uncivil perceptions? We need to test the external validity of the method in order to answer this two-pronged question.

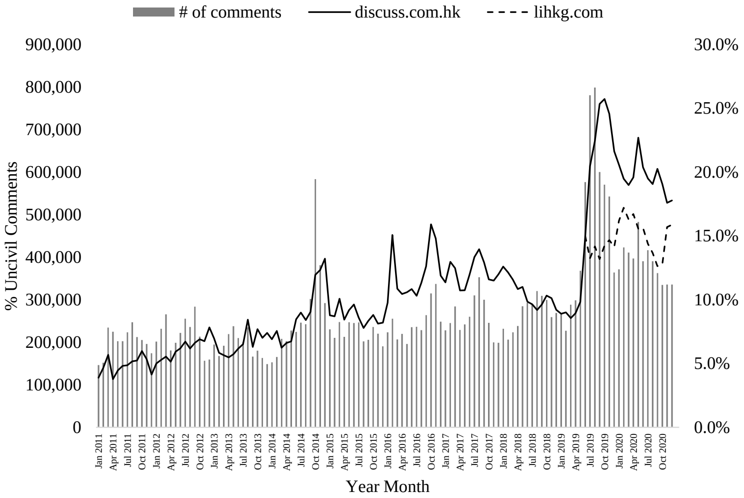

Figure 3: The percentages of uncivil comments in the two forums over time.

Incivility is pervasive on social media, particularly during times of social movement, arguably driven by the disinhibited and anonymous nature of social media (Cho & Kwon, 2015). Past scholarship has found that uncivil comments are more prevalent during discussions of more sensitive and controversial topics (Rossini, 2022). Previous literature has explored uncivil discourse within the context of the Hong Kong social movements. For instance, Chew (2023) conducted a study on how verbal violence was strategically and instrumentally utilized during Hong Kong's Anti-Extradition Movement. Chan et al. (2019) analyzed incivility and impolite speech on Facebook during the 2014 Umbrella Movement. They manually coded the contents of posts as impolite or uncivil and found that uncivil behavior is associated with cyberbalkanization (also see Lee et al., 2019). Likewise, Ng et al. (2022) studied how Hong Kongers use uncivil and supportive language and expressions to convey their negative emotions towards out-group Mainland Chinese on the HKGolden forum (also see Liang & Ng, 2022). They created a swearword dictionary with Cantonese profanities and foul language and discovered that Hong Kongers made more uncivil responses to posts about Mainland Chinese than to posts about Hong Kong.

Figure 3 illustrates the percentages of uncivil comments over time on LIHKG and DISCUSS, respectively. Several peaks on DISCUSS were associated with major political events in Hong Kong. The first peak in 2014 coincided with the large-scale Umbrella Movement. The second and third peaks in 2016 coincided with the Mong Kok "riot" and the legislative council oath-taking controversy. The last peak in 2019 reflected the social-emotional state during the Anti-ELAB Movement. The results demonstrate the face validity of our method since uncivil language is expected to be more prevalent during social movements. However, the use of uncivil language may also be associated with increased social media activity during political events in general. Figure 3 shows a moderate correlation between the total number of comments and the percentage of impolite comments after first-order differencing (*Spearman Rho* = .44, *p* <.001). Nonetheless, the correlation is not entirely perfect, as evidenced by the absence of a rise in the total number of comments during the Mong Kok "riot" in February 2016, despite an increase in the percentage of uncivil comments. Furthermore, the trend aligns with the public perception that impolite comments on LIHKG were more frequent before the Anti-ELAB Movement and remained stable during the Movement.

Survey experiment

In addition, to test whether the words in the dictionary are indeed considered uncivil by the general public, we conducted a representative online survey using Qualtrics's Hong Kong panel. Participants ($N = 822$) were sampled by age × gender according to Hong Kong's population (ages 18–65, Cantonese speakers); each participant was asked to rate 10 randomly assigned comments selected from a pool of 500 (of the 3,000) coded comments. Participants rated, on 7-point scales, their assigned comments on the four items of a perceived incivility scale developed by Kenski et al. (2020): uncivil–civil, impolite–polite, unnecessary–necessary, and disrespectful–respectful. A measure of perceived incivility was the average of those four items ($M = 3.90, SD = 1.75, Cronbach's\ alpha = 0.95$).

On average, comments with uncivil words were perceived as more uncivil than comments without uncivil words ($M_{uncivil}$ = 4.08 vs. M_{civil} = 3.52,

$t(5627.9) = 14.12, p < .001$). Furthermore, to avoid confounding impacts from different participants, a fixed-effect model conditioned on participants was estimated. The difference in perceived incivility between uncivil and civil comments was 0.60 ($SE = 0.03, p < .001$). The effect size was small, given that the measure used a 7-point range.

Given that each comment was independently coded by around 16 participants, we could consider the survey to be a form of crowdsourced coding. When perceived incivility by a participant was greater than four on the 7-point scale, a comment was labeled as uncivil; otherwise, it was labeled civil. To aggregate the incivility perception of different participants, we followed the rule of majority vote: if more than half of crow coders rated a comment as uncivil, we considered it uncivil. The agreement between the dictionary and crowdsourced coding was only 57.2%. Among the 341 comments with uncivil words, 191 (56.0%) were considered civil by the crowd coders.

Taken together, political incivility identified via the uncivil dictionary was associated with social reality (see Figure 3) and perceptions of incivility. However, the agreement between dictionary-based incivility and perceived incivility was weak. In general, people may consider the use of uncivil words on digital platforms as expected and even acceptable behavior that does not offend them in that context. This is understandable because perceptions of incivility in the same content depend on both individual and contextual factors, such as gender and partisanship (see Liang & Zhang, 2021; Massaro & Stryker, 2012). While the primary focus of this study is identifying uncivil content, predicting perceptions of incivility is also of interest. The crowd evaluation could then be considered the ground truth. Tables A1 & A2 in the **appendix** report the accuracy metrics of different methods. Textual features alone are not sufficient for predicting the perceptions of incivility (with accuracies around .60), though dictionary and human coders performed slightly better than the machine learning models. We also experimented with different thresholds to define incivility (i.e., with at least 40% or 35% of crowd coders perceived uncivil), and the patterns remained similar. The finding does not mean that dictionary outperforms machine learning in predicting incivility perceptions. Machine learning has the potential to incorporate individual and contextual factors, which could improve accuracy. However, our study found that machine learning, based solely on textual data, may not perform better than the dictionary-based method. Nevertheless, considering the generally low accuracies, dictionary-based methods may lack external validity and are not better than human coders in interpreting uncivil content.

Discussion

In summary, the present study proposes a dictionary enrichment method using word embeddings (Figure 1) and shows that the dictionary-based method could outperform supervised machine learning models (Table 1), including deep learning models (Table 2). The initial choice of seed words during the enrichment process can influence the final dictionary (Figure 2). However, even a small number of random seed words (e.g., $n = 10$) can generate a dictionary covering most uncivil incidences (94.5%) in the corpus. Furthermore, dictionary-based incivility coincides with major political events in Hong Kong (Figure 3) and is correlated with the perceptions of incivility by the public, which indicates a certain degree of external validity. Nevertheless, we found that incivility detected based on content features (comments containing vulgarity and name-calling) differs from incivility as perceived by the general public. In our case, more than half the vulgar or name-calling comments were not considered uncivil by our survey participants.

To be clear, we are not arguing that a dictionary-based method is universally better than the supervised machine learning approach when measuring communication concepts; rather, we demonstrate that dictionary construction remains valuable in some contexts, particularly in identifying uncivil Cantonese words in an online context. First, online comments are usually too short to be classified accurately by a simple machine. Short texts have inherent disadvantages like their lack of length, few features, and limited context that together provide a weak signal (Wang et al., 2017). In addition, as presented in Table 1, models based on word frequency outperformed the corresponding TF-IDF models. When text length varies greatly, it might be important to use TF-IDF measures (Rajaraman & Ullman, 2011). However, words are less likely to repeat within a single document in short texts like tweets. Thus, even a binary measure of presence is as informative as frequency count (Ikonomakis et al., 2005).

Second, deep learning algorithms require access to immense amounts of training data to achieve acceptable accuracy. A conventional approach to short-text classification is to associate short texts with existing knowledge bases (e.g., from search engines). Recent work has used pre-trained word embedding and deep learning for short text classification (e.g., Wang et al., 2017). The results in the present study suggest that deep learning with word embedding might not perform as well as the dictionary-based method. Increasing the size of the training dataset might improve the accuracy of deep learning algorithms, given that there are many parameters to be estimated.

However, increasing training cases also requires far more manual coding work, and the entire model-building task becomes less efficient.

Third, name-calling and vulgarity in Cantonese are usually based on keywords. Dictionary-based methods are especially effective in the identification and classification of keyword-based cases. From the conventional perspective, using vulgar language or name-calling is inappropriate, particularly when making public speeches, as it risks losing credibility and causing offense. Therefore, people come up with various euphemisms and minced oaths to replace off-limits words. However, this conventional perspective might not fully apply in online contexts. The survey results show that many comments with presumably uncivil words were actually perceived as civil—or at least as not uncivil—by the general public. Therefore, using those bad words does not necessarily elicit perceptions of incivility. Advanced machine learning models like BERT, with deep neural networks, might be more capable of dealing with this contextual problem. BERT learns contextual embeddings for words so that the same word could have different vector representations in different sentences (Devlin et al., 2019).

Limitations

Our proposed method also has certain limitations. First, it might be more accurate to describe this dictionary approach as an enrichment process, as the method is not perfect for constructing a holistic dictionary. Without experts' prior knowledge of the substantive areas, coverage of uncivil words is still high, but some rare but interesting words that are worth studying may be missed. Second, we used Word2Vec for word embeddings. Accuracy and coverage could be improved by using other recently developed models, such as fastText (Bojanowski et al., 2017; Joulin et al., 2016) or BERT (Devlin et al., 2019), which can model subwords and contextual variations. Third, the embeddings were trained based on forum data, and most comments on LIHKG are related to the 2019 Anti-ELAB Movement. Using data from a single online platform in a relatively narrow time period might not generalize well to diverse digital contexts. However, the proposed method can easily be replicated whenever more representative texts are available. Finally, the present study contrasted the dictionary method with supervised machine learning. However, previous studies have argued for combining dictionaries with supervised machine learning to improve accuracy (Dobbrick et al., 2022; Dun et al., 2021; Jakob et al., 2023). Future research should consider this approach. In this study, since the accuracy of our dictionary-based method is high enough (accuracy = 0.92), incorporating supervised learning would only increase the cost of manual coding in practice. Researchers need to make trade-offs between performance and cost, given that all automated content methods have their own advantages and disadvantages.

Supplementary materials

The dictionary, Word2Vec, and replication materials for tables and figures are available on **GitHub**. Appendix tables could be obtained here: https://github.com/rainfireliang/WORD-EMBEDDING-FOR-DICTIONARY-CONSTRUCTION/blob/main/appendix.pdf

Funding

This work was supported by Hong Kong Research Grants Council (GRF: 14600520).

References

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/10.1017/pan.2020.8

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Chan, C.-H., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., Van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27. https://doi.org/10.5117/CCR2021.1.001.CHAN

Chan, C.-H., Chow, C. S.-l., & Fu, K.-w. (2019). Echoslamming: How incivility interacts with cyberbalkanization on the social media in hong kong. *Asian Journal of Communication*, *29*(4), 307–327. https://doi.org/10.1080/01292986.2019.1624792

Chew, M. M.-t. (2023). The strategic and instrumental use of verbal violence by protesters: Political swearing in hong kong's anti-extradition movement. *Social Movement Studies*, 1–18. https://doi.org/10.1080/14742837.2023.2171384

Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, *51*, 363–372. https://doi.org/10.1016/j.chb.2015.04.046

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. *Proceedings of the fourth workshop on online abuse and harms*, 95–101. https://doi.org/10.18653/v1/2020.alw-1.12

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*, 512–515. https://doi.org/https://doi.org/10.1609/icwsm.v11i1.14955

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2022). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures*, *16*(4), 303–320. https://doi.org/10.1080/19312458.2021.1999913

Dun, L., Soroka, S., & Wlezien, C. (2021). Dictionaries, supervised learning, and media coverage of public policy. *Political Communication*, *38*(1), 140–158. https://doi.org/10.1080/10584609.2020.1763529

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. https://doi.org/10.1145/2858036.2858535

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Springer Nature.

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029. https://doi.org/10.1037/a0015141

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *2016*, 595–605. https://doi.org/10.18653/v1/D16-1057

Hovy, E. H. (2015). What are sentiment, affect, and emotion? applying the methodology of michael zock to sentiment analysis. *Language production, Cognition, and the Lexicon*, 13–24. https://doi.org/10.1007/978-3-319-08043-7_2

Ibrahim, M., Torki, M., & El-Makky, N. M. (2020). AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1881–1890. https://doi.org/10.18653/v1/2020.semeval-1.248

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966–974. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b0485fba23aabda526358f31cb5a382b66a08270

Jakob, J., Dobbrick, T., & Wessler, H. (2023). The integrative complexity of online user comments across different types of democracy and discussion arenas. *The International Journal of Press/Politics*, *28*(3), 580–600. https://doi.org/10.1177/19401612211044018

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. https://doi.org/https://doi.org/10.48550/arXiv.1612.03651Focustolearnmore

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. https://doi.org/10.1177/0093650217699933

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

Lai, R., & Winterstein, G. (2020). Cifu: A frequency lexicon of hong kong cantonese. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3069–3077. https://aclanthology.org/2020.lrec-1.375

Lee, F. L., Liang, H., & Tang, G. K. (2019). Online incivility, cyberbalkanization, and the dynamics of opinion polarization during and after a mass protest event. *International Journal of Communication*, *13*, 20.

Liang, H., & Ng, Y. M. M. (2022). The expression effects of uncivil disagreement: The mechanisms of cognitive dissonance and self-perception. *Human Communication Research*, *49*(3), 251–259. https://doi.org/10.1093/hcr/hqac032

Liang, H., & Zhang, X. (2021). Partisan bias of perceived incivility and its political consequences: Evidence from survey experiments in hong kong. *Journal of Communication*, *71*(3), 357–379. https://doi.org/10.1093/joc/jqab008

Luke, K. K., & Wong, M. L. (2015). The hong kong cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series*, (25), 312–333.

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*, 375–442.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 26–34. https://aclanthology.org/W10-0204

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Ng, Y.-L., Song, Y., & Huang, Y. (2022). Supportive and uncivil expressions in discussions on out-groups by in-group members in anonymous computer-mediated communication. *Telematics and Informatics*, *69*, 101785. https://doi.org/10.1016/j.tele.2022.101785

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. http://liwc.net/LIWC2007LanguageManual.pdf

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Rizos, G., Hemker, K., & Schuller, B. (2019). Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 991–1000. https://doi.org/10.1145/3357384.3358040

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49*(3), 399–425. https://doi.org/10.1177/0093650220921314

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for german user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Timm, J., & Barberá, P. (2019). Incivility begets incivility? understanding the contagion dynamics of uncivil conversations on social media.

Tong, C., Margolin, D., Chunara, R., Niederdeppe, J., Taylor, T., Dunbar, N., King, A. J., et al. (2022). Search term identification methods for computational health communication: Word embedding and network approach for health content on YouTube. *JMIR Medical Informatics*, *10*(8), e37862. https://doi.org/10.2196/37862

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 777–785. https://aclanthology.org/N10-1119

Wang, J., Wang, Z., Zhang, D., & Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, *350*. https://doi.org/https://doi.org/10.24963/ijcai.2017/406

Widmann, T., & Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2022.15

Wojcieszak, M., de Leeuw, S., Menchen-Trevino, E., Lee, S., Huang-Isherwood, K. M., & Weeks, B. (2023). No polarization from partisan news: Over-time evidence from trace data. *The International Journal of Press/Politics*, *28*(3), 601–626. https://doi.org/10.1177/19401612211047194

Yang, Y., & Eisenstein, J. (2015). Unsupervised multi-domain adaptation with feature embeddings. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 672–682. https://doi.org/10.3115/v1/N15-1069

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234