Advancing Smart Cities through Novel Social Media Text Analysis: A Case Study of Calgary

Mitra Mirshafiee, Ann Barcomb and Benjamin Tan
Department of Electrical and Software Engineering
University of Calgary
mitrasadat.mirshafie@ucalgary.ca, ann@barcomb.org, benjamin.tan1@ucalgary.ca

***Abstract*—In numerous cities, population expansion and technological advancements necessitate proactive modernization and integration of technology. However, the existing bureaucratic structure often hinders local officials' efforts to effectively address and monitor residents' needs and enhance the city accordingly. Understanding what people find important and useful can be inferred from their posts on social media. Twitter, as one of the most popular social media platforms, provides us with valuable data that, with the right tools and analysis, can provide insights into the performance of urban services and residents' perception of them. In this study, we used the city of Calgary as an exemplar to gather tweets and analyze topics relating to city development, urban planning, and minorities. Natural language processing (NLP) techniques were used and developed to preprocess stored tweets, classify the emotions, and identify the topics present in the dataset to eventually provide a set of topics with the prevalent emotion in that topic. We utilized a variety of methods to analyze the collected data. BERTopic for topic modeling and few-shot learning using Setfit for emotion analysis outperformed the others. Hence, we identify issues related to city development, senior citizens, taxes, and unemployment using these methods, and we demonstrate how delving into these analyses can improve urban planning.**

***Index Terms*—Emotion analysis, Topic modeling, Social computing, Social media, Smart society, Smart city, Analytics, Requirements gathering, Natural language processing**

I. Introduction

City modernization and urban planning have become some of the most critical issues that city authorities must tackle while faced with a rapidly growing city. As cities strive to become smarter and more responsive to residents' needs, understanding the preferences of urban residents is crucial. Social media platforms (e.g., Twitter) have emerged as a valuable source of real-time user-generated content, offering insights into residents' thoughts and opinions. Extracting meaningful information from the vast amount of social media text requires advanced techniques, and Natural Language Processing (NLP) provides a powerful solution.

To investigate how to harness relevant city-related information from social media, we perform a case study on Calgary, a city in Alberta, Canada, that is experiencing fast-paced modernization. The transformation is attracting many people – annual growth stands at about 3% – strengthening the economy and increasing vibrancy[1]. However, this growth is also putting pressure on the city's infrastructure, services, and management systems, making city modernization a significant challenge.

To achieve the benefits of city modernization and improve residents' satisfaction, authorities must continuously assess their actions through the eyes of residents to see whether the changes they plan to implement are aligned with the public perception of a modernized smart city. Specifically, the data received from residents in different forms might highlight perspectives and solutions the city is unaware of. It might also highlight tensions that should be addressed or indicate widespread misinformation/ignorance, which could be targeted through misinformation campaigns. Currently, the City of Calgary directly receives residents' feedback and requests by phone and via web submissions. Additional surveys[2] collected by the city employees are also another source for guiding which pathways are best for administrators to follow. To supplement existing manual approaches, we suggest a novel social media-based analytical method that also enriches the city's database with useful information that had not been previously unveiled. In this paper, our contribution is a framework that enables the following:

- Development and comparison of models for emotions analysis for city-related tweets from residents.
- Implementation and comparison of topic modeling techniques based on topic coherence and topic diversity.
- Generation of resident emotions, graphed over time for monitoring emotions.
- Generation of visualizations using the combination of topic modeling and emotion analysis results.

Overall, our goal in this work is to provide actionable insights for urban planners and policymakers to create more inclusive and responsive urban environments. This research lays the groundwork for further advancements in emotions and topic modeling, not just for Calgary but also for other smart cities worldwide.

II. Prior Related Work

Urban planners have hailed the introduction of smart city initiatives as a promising trend [1]. Smart city initiatives have won praise for their promise to use technology to boost cities' competitiveness and functionality, while also coming up with

[1]https://perma.cc/8LD2-2BNH

[2]https://www.calgary.ca/research/citizen-satisfaction.html?redirect=/citsat

innovative solutions to social issues such as poverty and social deprivation [2].

The documentation of current smart city initiatives points to several advantages of technology adoption [3]. A better understanding of residents' perceptions and engagement with these programs could provide important perspectives into public support and valuations regarding their needs. According to Hollands et al. [2], smart city authorities should include feedback from the public as a part of their planning and execution processes. Accordingly, studies that measure public opinion of smart cities have become more prevalent recently (e.g., [2], [4]–[7]). To date, examinations of people's opinions and responses have been limited and such research has never been done in regard to the rapidly expanding city of Calgary.

We chose Twitter as our target for this study as it is one of the most popular platforms and widely adopted among researchers for text analysis purposes [8]. Based on a report from the Twitter blog, more than 15 million Canadians use Twitter every month. That translates to 49.7% of Canada's online population. Of this group, close to half (44%) check in, consume and engage multiple times a day[3]. All of this demonstrates a high level of popularity and involvement, which indicates it is a great repository of public opinion.

III. Dataset

For the purpose of data collection, the Academic Research Product Track API was used to extract tweets between January 2020 and January 2023. We were able to use this tool to efficiently choose tweets that satisfied certain inclusion and exclusion criteria by picking the precise keywords we were looking for, e.g., "`(Calgary OR YYC) lang:en -fiwe -yeg -NowPlaying -ampcalgary`." The full query and criteria, longer than this line, are available on our GitHub page[4].

The user location was then used to filter out everyone who did not have "Calgary" in their location. In the end, we have 444,342 tweets which will be used for the examination of our models. Furthermore, for visualization of BERTopic, we have extended the inclusion criteria to view the most relevant tweets. During preprocessing, a set of model-specific cleaning steps were taken to prepare the inputs. For instance, using regex, the NLTK library, and spaCy packages, we removed hyperlinks, stop words, and irrelevant punctuation for LDA topic modeling. However, since semantic models are able to consider the context of a sentence as a whole, the removal of stop words was not needed for emotions analysis.

IV. Emotion Analysis

For Emotion analysis, we have investigated three approaches, discussed next. Each model needs its own cleaning procedure and training. The final model (Zero-shot technique), however, already has knowledge from prior training and is based on classification without any labels; hence no training is involved. We chose to classify four emotions – anger, joy, optimism, and sadness – based on Barbieri et al.'s evaluation framework (TweetEval) [9]. These will help us navigate the dominant emotion of people toward each topic revolving around the city.

A. Few-shot learning

Few-Shot Learning (FSL) has gained significant popularity in various tasks in NLP since the introduction of GPT-3 [10]. Brown et al.'s study demonstrated the improvement of task-agnostic few-shot performance by scaling up language models, surpassing earlier state-of-the-art fine-tuning approaches. In this study, we adopted SetFit, an efficient and prompt-free framework for few-shot fine-tuning of Sentence Transformers (ST) from Tunstall et al. [11].

The goal of this type of training (inspired by contrastive learning) is to minimize the distance between pairs of semantically similar sentences and maximize the distance between sentence pairs that are semantically distant. With the help of a distance-based loss function, such as cosine similarity, the embedding model is adjusted such that samples with various labels are separated in feature space. Using this method, the ST will be converted from a sentence encoder to a topic encoder, training a powerful, few-shot text classifier in a short time.

After choosing the model we want to use as the base for our algorithm ("sentence-transformers/all-mpnet-base-v2") we define the trainer configuration and run the model to be optimized. SetFit employs a two-stage training strategy. In the first stage, sentence pairs are used to fine-tune an ST on the input data, which may only contain a few labeled samples. Then, the rich embeddings (or sentence representations) created by the fine-tuned ST from the first stage are used to train a text classification head in the second stage.

B. Keyword approach

We also employed a method that includes the manual creation of a comprehensive list of specific words and emojis representing each emotion we aimed to classify. This method draws inspiration from Genc et al.'s [12] study, which used the keyword approach to classify tweets into news, opinions, deals, events, and private messages, achieving improved accuracy by leveraging eight fundamental features from tweets. Furthermore, Khattak et al. [13] integrated keyword searches into their pipeline for personalized tweet recommendations. Their proposed system was tested on a dataset of nearly 1 million tweets, which gained an accuracy of up to 96% in tweet classification. Building upon these previous studies, we implemented a Boolean technique to filter tweets using our curated list of words and emojis. This process allowed us to create a dataset where each record contained at least one of the designated keywords associated with the respective emotions we sought to identify. For instance, a tweet was categorized as "anger" if it contained a specific phrase or emoji denoting anger (e.g., "annoyance", 😡, "irritation"), while the presence of the "delight" or 😀 in a tweet qualified it as an expression of joy. To ensure our training dataset consists of tweets that accurately represented the emotions we are aiming to classify, we went back and forth in experimenting with

[3]https://blog.twitter.com/en_ca/topics/insights/2018/TwitterCanada_at_Dx3

[4]https://github.com/mitramir55/Advancing-smart-cities-with-nlp

different lists of words. After the construction of this list, we can build a labeled dataset that can be fed to a classifier. Our classifier was built using two separate language models, namely "cardiffnlp/twitter-roberta-base" and "vinai/berTweet-base", which are RoBERTa and BERT language models that were further trained on tweets. Based on our experiments, the RoBERTa-based model had greater accuracy. To the best of our knowledge, this is the first study to implement the combination of keyword searching and using semantic classification to analyze emotions surrounding citizens' opinions.

C. Zero-shot learning

Zero-shot learning (ZSL) is a method that generates predictions for new, unseen classes by using semantic information extracted from the model for the labels and documents. Therefore, this method is particularly valuable when working with no or limited labeled data and proves suitable for scenarios such as analyzing large social media records. This approach has also gained popularity across various research domains, including Alhoshan et al. [14], who proposed a novel approach that used ZSL for requirements classification. This study achieved an average recall and F-score of 82% on large volumes of unlabeled datasets.

In line with our study objectives, we adopted a comparative approach, leveraging sentence embeddings inspired by Reimers et al. [15]. This method has demonstrated superior performance in similar classification tasks. Furthermore, we follow the semantic textual similarity methodology to the one they outlined to build upon their established framework.

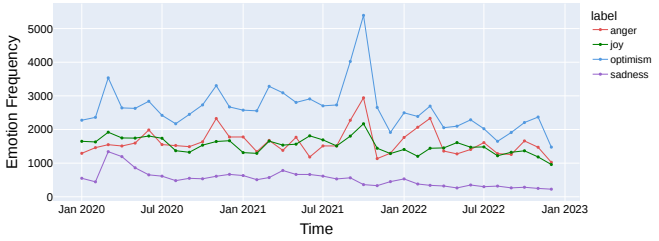

Fig. 1. Emotions distribution of all tweets through time

V. Topic modeling

For the purpose of identifying the subjects people talk about, we implemented three topic modeling approaches. Our first model, Latent Dirichlet Allocation (LDA), is a traditional and commonly used algorithm in the field of NLP. We compare this statistical model with two other recently developed models, Combined Topic modeling and BERTopic. We will see how semantic and textual knowledge the latter two utilize can affect model performance and produce interpretable topics.

A. Latent Dirichlet Allocation (LDA)

LDA is a commonly used topic modeling algorithm introduced by David Blei et al. in 2003 [16]. LDA is a generative probabilistic model (a three-level hierarchical Bayesian model) that enables data observations to be explained by latent variables, which reveal why particular documents are related to a set of subjects. The fundamental concept is that each text document is represented as a random mixture of latent topics [17], where each subject is defined by a distribution over the words that are present in the document.

B. Combined Topic Model

Contextualized Topic Models (CTM) are a family of topic models that facilitate topic modeling by using pre-trained representations of language (such as BERT). Neural topic models' topic coherence significantly increases when contextual information is added [18]. Bianchi et al. created Combined Topic Model (Combined TM) in an effort to integrate transferred knowledge into previous topic modeling techniques [18]. Overall, the SBERT embedded representations [15] are the main foundations for their model, which allow the rapid production of contextualized document embeddings. This approach has proven effective for monolingual datasets [18] and social media analysis [19].

C. BERTopic

In 2022, Grootendorst devised BERTopic [20], another method for using pre-trained language models to incorporate semantic information into topic modeling. This method utilizes clustering, hierarchical algorithms, and transfer learning. It draws on Top2Vec's techniques [21] which means their algorithmic structures are somewhat similar. However, the use of the class-based term frequency inverse document frequency (c-TF-IDF) algorithm, which evaluates the relative weights of words within a cluster and generates term representation, is one of the novelties of this method which implies that a phrase is more reflective of its topic the higher its TFIDF value [22].

In previous research, Sia et al. [16] showed the validity of clustering embeddings with centroid-based methodologies for representing topics. By embedding words and looking for those that are near a cluster's centroid, topic representations are retrieved from these clustered embeddings. In this algorithm, topic representations are made by locating terms near a cluster's centroid after documents have been grouped. Notably, while the clusters are produced using HDBSCAN (a hierarchical clustering algorithm), the topic representations are produced using a centroid-based viewpoint.

In summary, in BERTopic, Grootendorst builds a topic model to produce coherent topic representations by utilizing clustering methods and a class-based variant of TF-IDF. To get document-level information, he first builds document embeddings using a language model that has already been pre-trained on millions of documents. After reducing the dimensionality of document embeddings, he generates clusters of documents that are linked semantically and represent specific topics. Third, a class-based variant of TF-IDF is created to extract words representing the topics with a centroid-based perspective. These three separate phases make it possible to create a versatile topic model that can automatically identify topics in a dataset. Given the promising performance from prior work on various

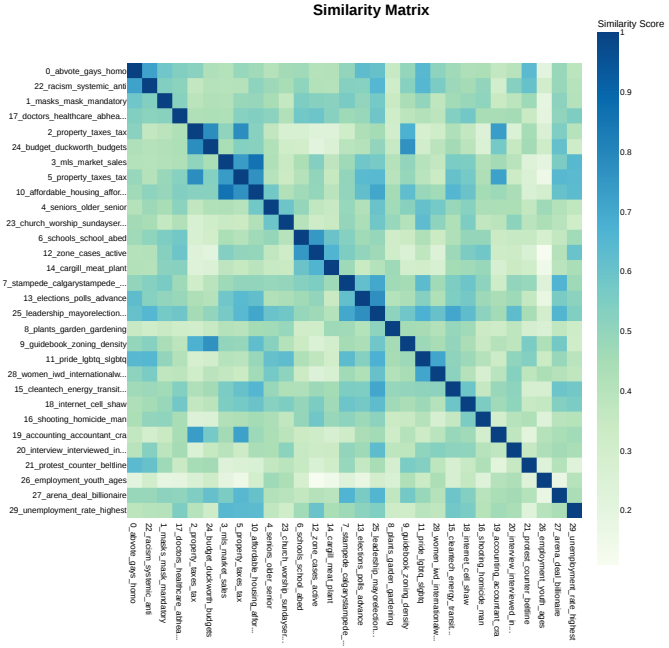

Fig. 2. Topics derived from BERTopic model and their relation to one another

TABLE I
RANKING MODELS' ACCURACIES

Method	Accuracy
Few-shot learning	0.83
Keyword approach + RoBERTa	0.74
Keyword approach + BERT	0.67
Zero-shot learning	0.61

news datasets [20], press coverage, and public perception [23], we explore this technique in our problem domain.

VI. EVALUATION

To test our emotion analysis methods, we randomly chose 160 records and manually labeled them as having the following emotions: "anger", "joy", "optimism", and "sadness." Eighty percent of this dataset has been selected for training our Few-shot model, and the rest is selected for the comparison of models. The choice to label 160 records and evaluate their performance was made to explore these models in contexts with few labeled samples for training.

In comparison to other techniques, few-shot learning produces the best accuracy (Table I). As a result, we select this model and produce a fully labeled dataset.

For Topic modeling, two commonly used metrics to evaluate models are topic coherence and topic diversity. Using normalized pointwise mutual information, the topic coherence of each topic model was evaluated. It has been demonstrated that this coherence metric can perform reasonably well to mimic human judgment [24]. The range of the measure is [-1, 1], with 1 denoting a perfect association. Subject diversity is

TABLE II
TOPIC MODELING TOPIC COHERENCE AND DIVERSITY

Model	Topic diversity			Topic Coherence		
	Number of topics			Number of topics		
	10	20	100	10	20	100
LDA	0.76	0.80	0.86	0.007	0.017	-0.01
CTM	1.0	0.99	0.51	0.14	0.13	0.15
BERTopic	0.97	0.98	0.93	0.12	0.19	0.11

the proportion of unique words across all themes [25]. The scale goes from [0, 1], with 0 denoting repetitive themes and 1 denoting topics with more variety. We employed the Octis topic modeling library [26] to implement these two measurements.

We ran our models based on three topic sizes (10, 20, 100). We should mention that BERTopic can determine the topics automatically since it uses a hierarchical algorithm. However, for a fair comparison, we chose to compare the models based on these three topic sizes.

We present our results in Table II. As we can see, BERTopic can serve as the best model for topic modeling techniques for this sort of data, based on taking topic diversity and coherence. The BERTopic package also has the advantage of having easy-to-use functions that help us visualize topics and their relations.

After comparing and determining BERTopic as our chosen model, we run it on our dataset and let it determine the present topics automatically. In Figure 2, we can see the top 30 most frequent topics and their similarity, based on the similarity of their representative word group. For instance, we can see certain topics revolving around the following subjects appear closer in vector space: tax, budget, and property (2, 5, 10); election and leadership (13 and 25); racism and homophobia (0, 22); LGBTQ+ and women's rights (11, 28).

We note that even though in the data cleaning and collection phase, we eliminated as many unrelated tweets as possible by defining exclusion and inclusion criteria, there may have been tweets with unrelated topics. Fortunately, BERTopic can see these as either noise (named Topic -1, which is not included in the visualizations) or put them together in one topic.

We bring together the results from topic modeling with the whole classified dataset to obtain insight into how people respond to various topics. In Figure 1, we see the dominant emotion is optimism. Anger and joy are closely related in frequency, and sadness comes last. We should keep in mind that after looking at tweets classified as optimism, we can see that some of them are news about a specific topic and tend to have a neutral tone. Hence, we chose to focus more on the other emotions while analyzing the results.

There are numerous topics that can be explored, but here we highlight a few. For instance, topic 9 in Figure 6 revolves around the city's plans for development. It has a sudden spike in May 2021 and when we look at comments in that period, we see people's complaints about the newly introduced plans:

"$200 Million to remodel downtown is short sighted. Unless we change the narrative &; stop throwing money at a problem &; spend some time, energy &; focus on creating a user &; business friendly environment at #yyccc#yyc"

Topic number 5 revolves around house prices and taxes in Calgary, which has been one of the main concerns of residents. This topic was associated with the following words: property, taxes, tax, increases, lowest, increase, homeowners, rate, hike, and residential. The most prevalent emotions are projected to be anger and optimism. After looking at the data, we see angry comments like "@USER This isnt good, real estate prices are already divorced from disposable income. Calgary is going to end up like Toronto and Van, residents priced out not by people from Canada but foreign investors, this should be a big concern to your party. AB needs a strong foreign buyers tax." Another example is from the tweets classified as optimism: "Calgary city council prepares to make budget adjustments to try to keep property taxes flat in 2022 #yyc".

Another topic is one revolving around senior residents (Figure 5). We see that joy and optimism are the dominant emotions that are directed toward the lives of the elderly and their condition in Calgary. Some of the tweets are as follows: "Thanks for the opportunity. Our Calgary Coordinated Response to missing Seniors is such a great example of collaboration-locally, provincially, nationally and even internationally." (optimism) and "@USER Thank you for your support of seniors in our community!" (joy) In topic 29, we have residents' requests and concerns regarding unemployment in Calgary: "@USER, Calgary's unemployment rate is 7.2%, and represents the WORST rate of any major city in Canada. Pair this with Edmonton (6.9%) unemployment and we have the 2 major problems. Will you finally admit your 33% corporate tax cut failed to create jobs for all Albertans?" and "@USER So here in Calgary the unemployment rate is what 15%. Please outline how bringing in more people will lower this rate? These new Canadians are initially a huge drag as they are getting social benefits and paying little taxes. Liberal math is really hard" Another interesting topic to look at is the one related to a city-wide event in Calgary called Stampede. This event occurs in July each year, and we can see a sudden spike in joy during this period. However, it is also important to note that anger tweets are present as well. For instance, in 2021, residents have expressed their worries about COVID-19 protocols and the spread of disease if the city holds the event: "That show kids &; schools are a conduit for spread of C-19. You've stopped testing for variants. And now? You plan to have the Calgary Stampede go ahead. Will you open it up to the high volume of tourists that normally attend? What about the competitors?"

As we've seen in this discussion, the combination of our suggested techniques (Figure 3) allows us to learn about topics such as taxation, seniors, and unemployment, which has substantial potential in assisting government officials in their decision-making process.

VII. Limitations

We recognize that not all residents of Calgary may have their location specified or made public on their profile, which could affect the diversity of our data. In addition, we demonstrate the capability of our emotions analysis models to identify accurately with little to no labeled data. However, the optimal performance of the few-shot learning model may also be compromised by the small dataset. Finally, we appreciate the usefulness of the keywords we selected for our keyword method and for collecting the dataset. While there may have been others that we missed that could have impacted the inclusivity of our results.

Furthermore, we acknowledge that there might be ethical concerns with the use of data derived from social media. Governments must be careful not to rely solely on these results when making decisions because, occasionally, special-interest groups may use bots to manipulate users on social media, and the objectives of these groups go against the government's legal and moral obligations to all residents.

VIII. Conclusion

In this paper, we explored the application of emotion analysis and topic modeling in analyzing social media text and its role in advancing smart cities. Among various techniques, few-shot learning demonstrated the highest accuracy and was chosen for analysis. By using this model, we created a fully labeled dataset, enabling us to understand dominant emotions expressed in social media posts. Combining topic modeling with the classified dataset revealed valuable insights into how Calgarians respond to different topics. We saw how residents are voicing their concerns, complaints, and suggestions through these comments and that knowing about these concerns may be very helpful to the local authorities as they decide what to do next. In conclusion, this study demonstrates an efficient way of text classification and topic modeling of social media text which can be quite useful for cities' authorities.

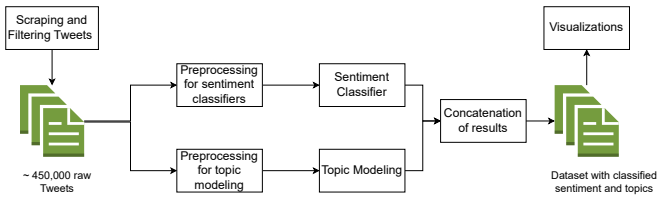

Fig. 3. Overall approach

References

[1] C. Harrison *et al.*, "Foundations for Smarter Cities," *IBM Journal of Research and Development*, vol. 54, no. 4, pp. 1–16, Jul. 2010, conference Name: IBM Journal of Research and Development.

[2] R. G. Hollands, "Will the real smart city please stand up?: Intelligent, progressive or entrepreneurial?" *City*, vol. 12, no. 3, pp. 303–320, Dec. 2008.

[3] A. Buttazzoni, M. Veenhof, and L. Minaker, "Smart City and High-Tech Urban Interventions Targeting Human Health: An Equity-Focused Systematic Review," *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2325, Mar. 2020.

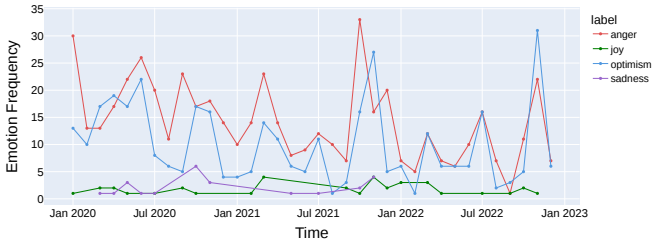

Fig. 4. Topic related to Tax and Property - associated words: property, taxes, tax, increases, lowest, increase, homeowners, rate, hike, residential

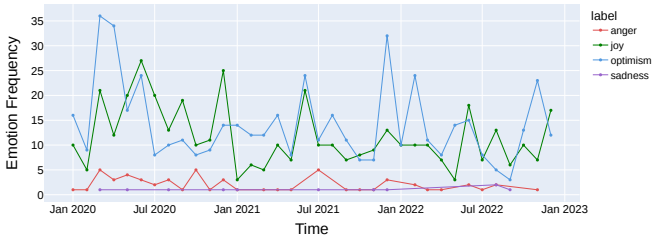

Fig. 5. Topic related to Senior residents - associated words: seniors, older, senior, adults, seniorsweek, aging, cards, yycseniors, seniorcare, activeaging

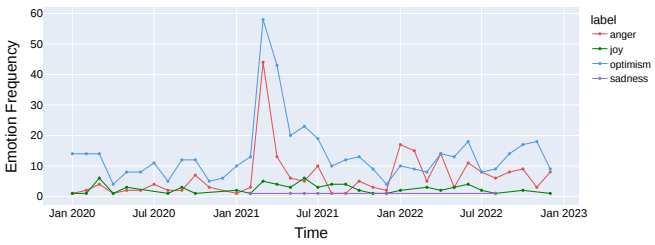

Fig. 6. Topic related to city development - associated words: guidebook, zoning, density, downtown, communities, area, planning, land, development

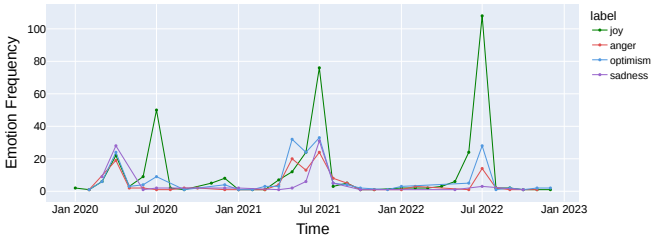

Fig. 7. Topic related to Stampede - associated words: stampede, calgarystampede, yahoo, rodeo, cancelled, breakfast, grounds, parade, spirit, pancake

[4] A. Adikari and D. Alahakoon, "Understanding Citizens' Emotional Pulse in a Smart City Using Artificial Intelligence," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2743–2751, Apr. 2021.

[5] S. El hilali and A. Azougagh, "A netnographic research on citizen's perception of a future smart city," *Cities*, vol. 115, p. 103233, Aug. 2021.

[6] A. Georgiadis, P. Christodoulou, and Z. Zinonos, "Citizens' Perception of Smart Cities: A Case Study," *Applied Sciences*, vol. 11, no. 6, p. 2517, Mar. 2021.

[7] N. Habbat, H. Anoun, and L. Hassouni, "Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets," in *Innovations in Smart Cities Applications Volume 4*, ser. Lecture Notes in Networks and Systems, M. Ben Ahmed *et al.*, Eds. Cham: Springer International Publishing, 2021, pp. 147–161.

[8] A. Keramatfar and H. Amirkhani, "Bibliometrics of sentiment analysis literature," *Journal of Information Science*, vol. 45, no. 1, pp. 3–15, Feb. 2019.

[9] F. Barbieri *et al.*, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650.

[10] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33. Currtoan Associates, Inc., 2020, pp. 1877–1901.

[11] L. Tunstall *et al.*, "Efficient Few-Shot Learning Without Prompts," Sep. 2022, arXiv:2209.11055 [cs].

[12] Y. Genc, Y. Sakamoto, and J. V. Nickerson, "Discovering context: classifying tweets through a semantic transform based on wikipedia," in *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems: 6th International Conference, FAC 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011. Proceedings 6*. Springer, 2011, pp. 484–492.

[13] A. M. Khattak *et al.*, "Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation," *Complexity*, vol. 2020, p. 8892552, Dec. 2020, publisher: Hindawi.

[14] W. Alhoshan *et al.*, "A Zero-Shot Learning Approach to Classifying Requirements: A Preliminary Study," in *Requirements Engineering: Foundation for Software Quality*, V. Gervasi and A. Vogelsang, Eds. Cham: Springer International Publishing, 2022, vol. 13216, pp. 52–59, series Title: Lecture Notes in Computer Science.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, arXiv:1908.10084 [cs].

[16] S. Sia, A. Dalmia, and S. J. Mielke, "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!" Oct. 2020, arXiv:2004.14914 [cs].

[17] B. Chen, "Latent topic modelling of word co-occurence information for spoken document retrieval," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 3961–3964, iSSN: 2379-190X.

[18] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence," Jun. 2021, arXiv:2004.03974 [cs].

[19] H. Alhuzali, T. Zhang, and S. Ananiadou, "Emotions and topics expressed on twitter during the covid-19 pandemic in the united kingdom: Comparative geolocation and text mining analysis," *J Med Internet Res*, vol. 24, no. 10, p. e40323, Oct 2022.

[20] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, arXiv:2203.05794 [cs].

[21] D. Angelov, "Top2Vec: Distributed Representations of Topics," Aug. 2020, arXiv:2008.09470 [cs, stat].

[22] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*, 1st ed. Cambridge University Press, Oct. 2011.

[23] G. Hristova and N. Netov, "Media coverage and public perception of distance learning during the covid-19 pandemic: A topic modeling approach based on bertopic," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 2259–2264.

[24] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539.

[25] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020, place: Cambridge, MA Publisher: MIT Press.

[26] S. Terragni *et al.*, "OCTIS: Comparing and optimizing topic models is simple!" in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Apr. 2021, pp. 263–270.