Interactive Sentiment Analysis and Clustering of Hiking Trails in Nova Scotia

Razia Choudhury Helia Rahimi Noah Barrett

***Abstract*—Nova Scotia is globally renowned for the scenic views found on the relatively large number of hiking trails in the province. The growing availability of mobile apps developed for the purpose of recording and sharing trail information provided us an interesting opportunity to gain insight into understanding individual's perspective on the included trails. In this report, we document the methods involved in building a graphic tool that encompasses explanatory and exploratory analysis of the trail data recovered from one of the trail guide apps of Nova Scotia. This visual interpretation engages several chart types and produces valuable insights that are documented in the report.**

***Index Terms*—Clustering, Hiking Trails, Sentiment Analysis**

I. Introduction

In terms of hiking trail ranking with the entirety of Canada: Skyline trail in Cape Breton National Park ranked 2nd in [1] for best hike in Canada, Kejimikujik National park ranked 3rd in [2] and both of these trails also sequentially ranked first and second for best hikes in [3]. These ratings only speak to two of the hundreds of trails in Nova Scotia, a large majority being equally as beautiful. One of the most popular services for recording and sharing trail information alltrails.com [4] was scraped to take advantage of the exploratory opportunity the trail guide apps entail.

The project aimed to investigate geographical relationships with various interesting properties of hikers' experiences on the trail. By combining metadata about the trail such as location, length, elevation gain and tags with real written reviews of the trail, a novel perspective on hikers' experiences was discovered. The project focused on a Visual Analytics perspective, jointly leveraging machine learning technologies such as Sentiment Analysis, Natural Language Processing, LDA Topic Modelling and Clustering with a powerful human-centric approach to data visualization.

The project findings would benefit hike-lovers to comprehend the condition of a specific trail from a hiker's perspective, find similar trails of interest, plus support decision-making. Also, this project could be of utter importance to the Government Departments/ Respective authorities to improve the trail conditions, thereby encouraging a spectacular hiking experience for a larger audience.

II. Problem Formulation

The problem under consideration is analysis of people's opinions on hiking trails of Nova Scotia to provide a comprehensive understanding of the included hikes' condition and encourage fact-based decision making.

A. Hiking Trails Dataset

As mentioned, the dataset for this project was collected by scraping the website alltrails.com [4] inclusive of all the hiking trails in Nova Scotia. This data set contains the metadata and reviews of all 713 hiking trails documented in Nova Scotia. Meta data for each trail was collected and stored in a csv format. Each trail properties (length, elevation, tags and written reviews) were uniquely identified using the trail_id.

B. Analysis Questions

- Which are the most liked trails in Nova Scotia?
- Which hike trails are negatively reviewed by people?
- What are the most prominent words for the positively liked trails?
- What negative words have been used for the included trails?
- What is the probability of words occurring in the negatively reviewed trails?
- Which trails are similar to the selected trail?

III. Methodology

A. Sentiment Analysis

To assess peoples' perspective of hiking trails, Sentiment analysis [5] (also known as opinion mining), a form of natural language processing technique that can be used to detect public opinions (e.g., positive, or negative, polarized, etc.) and interests within text or a whole document was used. Each trail's written reviews were considered as a single document leading to a data corpus of 713 documents (trails). Using python Library, TextBlob [6], a sentiment analysis model was created to detect polarity (tells how positive or negative a text is) and subjectivity - evaluating opinions and feelings in regards to the central subject. The polarity metric was used to evaluate the associated sentiments of the reviews. If the polarity score obtained using TextBlob was less than 0, the trail was marked as negatively reviewed, polarity equal to 0 identified neutrally reviewed trails and polarity greater than 0 indicated positively reviewed trails.

B. Text Analysis using Natural Language Processing

Text analysis was performed to transform unstructured text in reviews in a suitable form. Data preprocessing methods such as converting strings to text, maintaining case uniformity and

removing missing values, incorrectly formatted data, redundant words (e.g. hike, walk, trails, etc), stopwords, punctuations, graphical entries, were employed to extract meaningful information from text.

To further improve accuracy, Natural language processing (NLP)[7] techniques were employed. NLP, a branch of Artificial Intelligence (AI), helps the machine to understand, interpret and manipulate human language while responding in the same manner. Using python library, NLTK (processes human language data), the following NLP methods were applied on the cleaned datasets.

1) Tokenization [8], process of breaking down text into word units – a set of tokens using stop-words and lexical dictionary.
2) Contraction map was created to convert words like where've to where have, I didn't to I did not, etc.
3) Part-of-Speech(POS) tagging using TreeBank tags and Wordnet Lemmatization were implemented. Part-of-speech taggers typically take a sequence of words (i.e. a sentence) as input, and provide a list of tuples as output, where each word is associated with the related tag – nouns, verbs, adverbs,etc. Part-of-speech tagging [9] provided the contextual information (understanding the role of a word in a sentence) that a lemmatizer needs to choose the appropriate lemma (different inflected forms of a word).
4) After all these operations were performed, the tokenized reviews were stitched back together to visualize the most frequent words used in the positive and negative reviews. WordCloud is a visualization wherein the most frequent words appear in a generous size, and the less frequent words appear in smaller sizes, was used to see how well the given sentiments are distributed across the data corpus of words.

C. Topic Modelling using Latent Dirichlet Allocation (LDA)

LDA modelling [10] is a generative probabilistic model used to explain similarity between words in a document. For this project, we assessed the topic distribution of the negatively reviewed trails and identified the high probability of words in each topic using the most optimal LDA model. The following steps were used to achieve this:

- Filtered the words occurring in more than 90 percent and less than 10 percent of the negative reviews.
- Applied the term frequency-inverse document frequency (TF-IDF) transformation to give more weightage to unique words in the data corpus comprising negative reviews. TF-IDF [11] works by penalizing the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus but appear in good numbers in few documents.
- GridSearch and parameter tuning using a chosen list of number of topics and learning decay parameters were used to control the learning rate.
- The different topic models were evaluated using Log Likelihood and perplexity scores and further compared to choose the optimized model. Higher the Log Likelihood score the better is the model.

D. Clustering

1) Feature selection: We have used both numerical and categorical features to cluster trails. Numerical features contain the length and elevation of the trails, and categorical features consist of their types and a list of tags describing their characteristics. Our data includes a total of 53 unique tags with different frequencies, as shown in Figure 1. In our first approach, all features are utilized by considering each tag as a feature; besides, one hot encoder is applied to the type feature resulting in three more attributes (i.e. loop, out back, point to point) in addition to numerical ones. In our second approach for feature selection, the most frequent tags are employed along with the three types and numerical features.

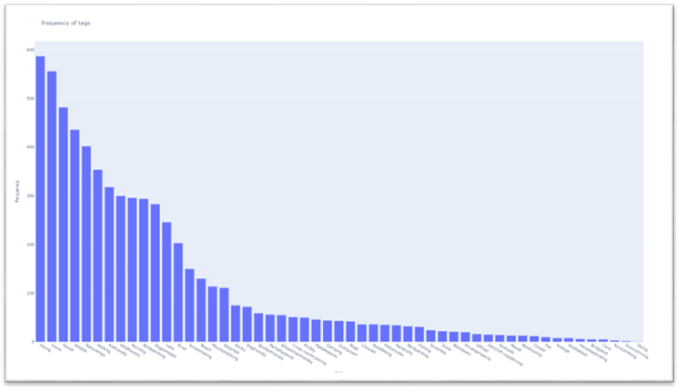


Fig. 1. Tags Frequency

2) Clustering Model: Tags, that comprise most of our features and represent our trails, can have a hierarchical manner. Hence, we practiced both partitional and hierarchical clustering in our project. We adopted Agglomerative clustering which has a bottom up approach, denoting that each trail is a cluster itself and clusters are successively merged together based on the linkage metric (i.e. ward). Also, we adopted KMeans and KMedoids clustering methods, which are partitional approaches. KMeans is an iterative clustering algorithm that aims to partition dataset into K non-overlapping subgroups where data points in each cluster are very similar to each other while also very different from data points in other sets. KMedoids which works similarly, aims to minimize the sum of distances between each point and the medoid- a data point with the least total distance to the other members- of its cluster. The advantage of KMedoids is that it can be more robust to noise and outliers in comparison with the KMeans. Figure 2 represents each methods and parameters we used in clustering the trails along with the silhouette score of each model.

Modelling Approach	Modeling Parameters	Number of Clusters	Silhouette Score	Feature Selection Approach
Agglomerative clustering	**Affinity*:** Euclidean **Linkage**:** Ward	4	0.829	All Tags
			0.78	Most frequent tags (freq>200)
		6	0.828	All Tags
			0.828	Most frequent tags (freq>200)
KMeans	**Init***:** k-means++	4	0.829	All Tags
			0.807	Most frequent tags (freq>200)
		6	0.828	All Tags
			0.828	Most frequent tags (freq>200)
KMedoids	**Init:** k-medoid++	4	Different values Best= 0.829 Worst= 0.675	All Tags

* **Affinity:** metric for computing the linkage.
** **Linkage:** determines which distance to use between clusters.
*** **Init:** method for initialization of centroids and medoids.

Fig. 2. Description of different clustering approaches used in this project.

Considering that the initializing points for each medoids affects the result of clustering, we discovered different outputs for KMedoids. However, the best result is the same as KMeans. Given the inconsiderable discrepancy between scores for all methods, we select KMeans as our clustering model due to its relatively simple implementation, guaranteed convergence, and its ability to generalized to clusters of different shapes. Figure 2 presents the relationship between number of clusters and silhouette score for KMeans using different feature selection approaches. This shows that the best silhouette score is achieved with 4 number of clusters.

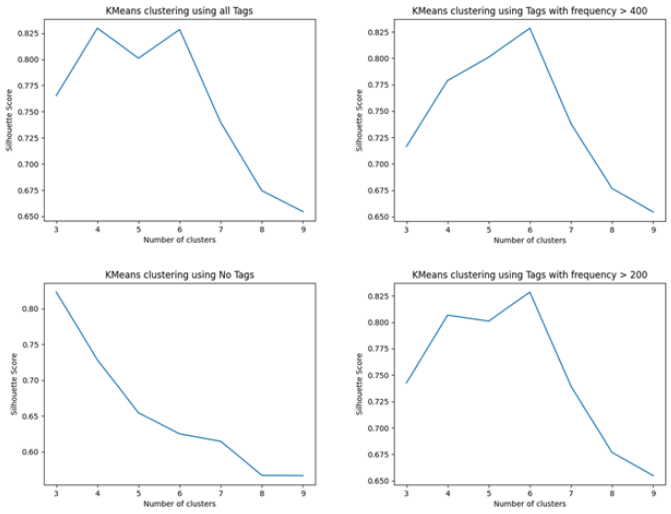


Fig. 3. Relation of silhouette score and number of clusters in KMeans.

3) Results: The final cluster is modelled by KMeans 4-cluster, with KMeans++ as the initializing method, using all tags, three types and two numerical features of length and elevation. The silhouette score is 0.83 indicating that our clusters are separate and distinguishable. Figure 3 represents the relation between numerical features (i.e. length and elevation) and types of trails with respect to clusters.

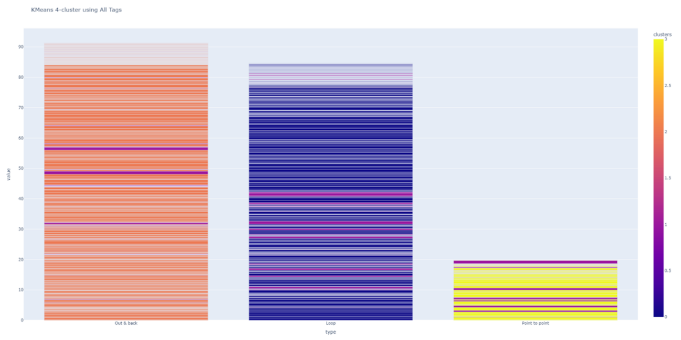


Fig. 4. Result of clustering. Relation of types and value= elevation + length.

As shown in Figure 4, three clusters (2, 0, and 3) have the characteristics of each types of trails (out back, loop and point to point respectively) and the fourth cluster (1) is scattered in other types. It also demonstrates that most of the trails in cluster 1 have small values of length and elevation.

IV. Dashboard and Interactivity

The developed application has two main views, the global view and local view. The global view is intended to serve as a aggregate visualization of many selected trails. In the global view, the user can select many trails and visualize the sentiment of the trails at a global level. The local view is intended to give users an individualized view of the sentiment of the trail and compare with similar trails based on the clustering results.

A. Global View

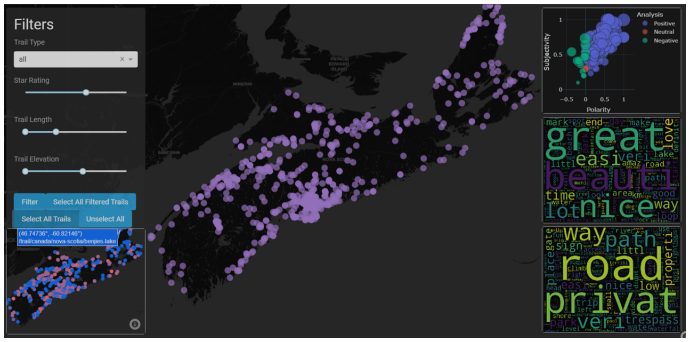


Fig. 5. Global view of Dashboard, users are able to visualize the sentiment of selected trails, as well as visualize a global view of clusters present in filtered trails.

B. Main Map

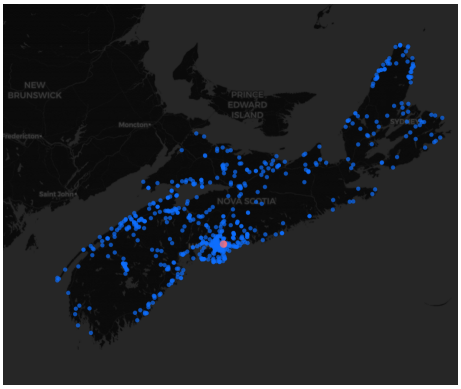


Fig. 6. Main Map in dashboard, used for selecting new trails and visualizing location of already selected trails.

In the centre of the global view of the dashboard seen in figure 6, a large map of Nova Scotia is used as mechanism for selection of new trails and puts geographical context to those that have already been selected.

1) Filtering Trails: During the process of searching for trails to perform sentiment analysis on, filtering through different characteristics of the trail such as elevation, and average rating can speed up the process. In the Upper Left of the global view of the dashboard as shown in figure 7, this functionality is provided.

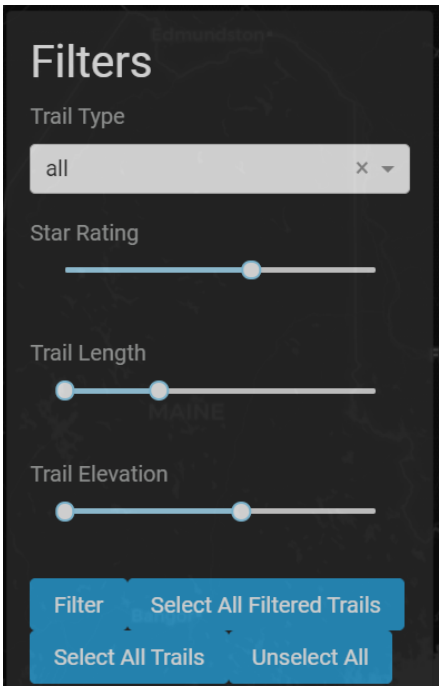
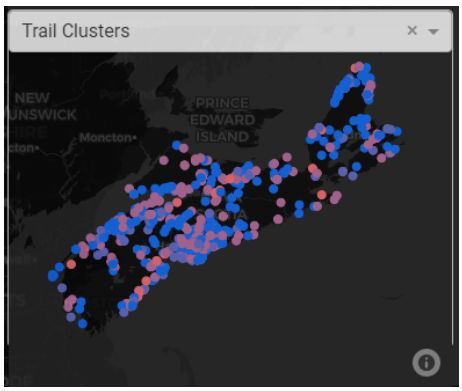
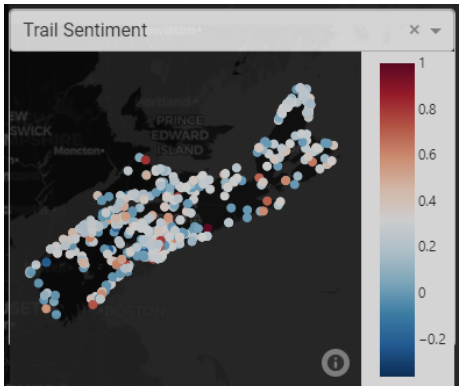


Fig. 7. Filter form to allow users to easily select large groups of trails with specific properties

The filter form pictured in figure 7 has several different functionalities. First, the user can select the type of trail they are interested in, filter the average star rating of the trail, trail length and elevation, when the desired choices are selected, the user can get a global view of the filtered trails by clicking the Filter button. This button will filter the trails viewed on the main map shown in figure 6. Additionally, the user has the option to select all the filtered trails, allowing for speedy selections of multiple trails to visualize their sentiment analysis. In the bottom row of buttons, the users have the option to select all the trails, providing a global sentiment analysis of all of the trails in Nova Scotia, as well as a unselect all button that unselects all trails that are currently selected.



(a) Selected Global Cluster View



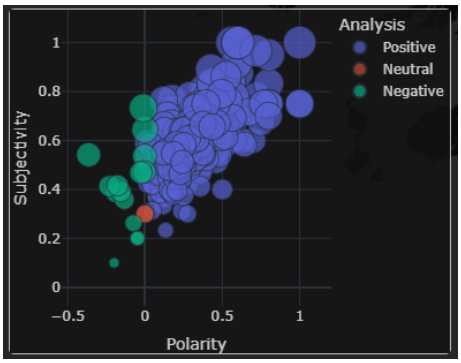
(b) Selected Global Sentiment View

Fig. 8. View of selected trail clusters and sentiment

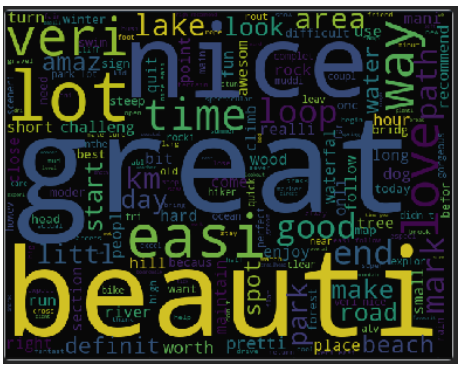
2) Global view of Selected trail Clusters and trail sentiment: On the bottom left side of the global view of the dashboard seen in figure 8, all of the selected trails are visualized, the user is able to toggle between visualizing the trail's sentiment and the trail's associated cluster. For the cluster view, the color of each point represents the cluster that they belong to. For the sentiment view, the color represents the polarity of the trail, allowing to visualize the trails sentiment relation to each other.

3) Sentiment Analysis: On the right side of the global view of the dashboard seen in figure 9, there are three sentiment analysis visualizations. These visualizations aim to streamline the process of interpreting how hikers *feel* about the selected trails. In figure 9 (a), the users are able to visualize the

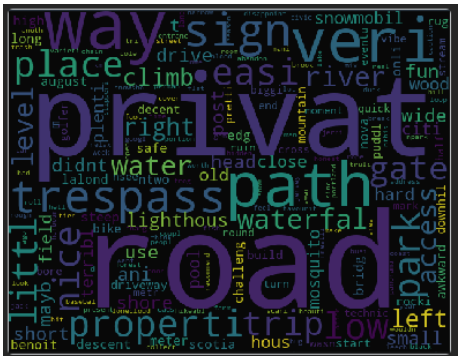
sentiment against the polarity, a useful measure in sentiment analysis. figures 9 (b) and (c) allow the user to interpret the words that attributed to the results in 9 (a).



(a) Sentiment Scatter Plot



(b) Positive Sentiment Word Cloud



(c) Negative Sentiment Word Cloud

Fig. 9. Global Sentiment Analysis of selected trails

C. Local View

The local view is intended to give the user a better understanding of a specific trail, it provides visualizations of the sentiment analysis of the specific trail with respect to trails in the same cluster, a general description of the trail, average rating, some basic stats on the trail and a map view showing the location of all similar trails. As can be seen on the top of figure 10, users can choose to add a trail to their selected trails by clicking the select button.

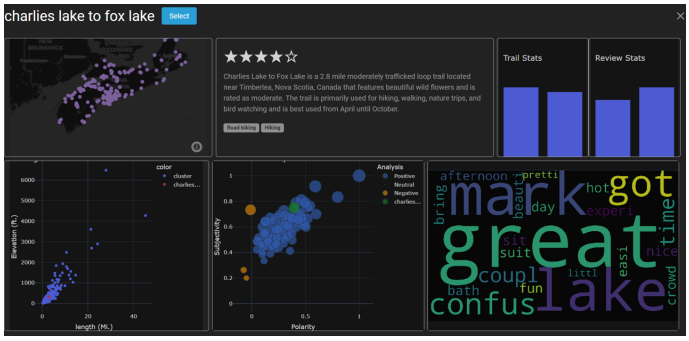


Fig. 10. Local view of Dashboard, users are able to visualize individualized information of a selected trail.

1) Similar Trail's Locations: In the top right corner of figure 10, a map showing the locations of all similar trails to selected trail is visualized, this gives the user a geographic understanding of the selected trail with respect to similar trails in Nova Scotia. This visualization is pictured in figure IV-C1.

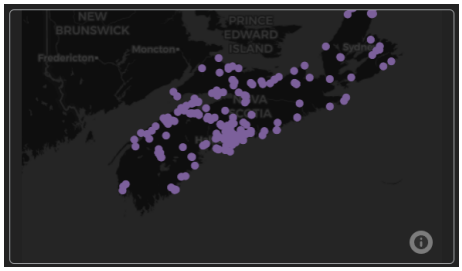
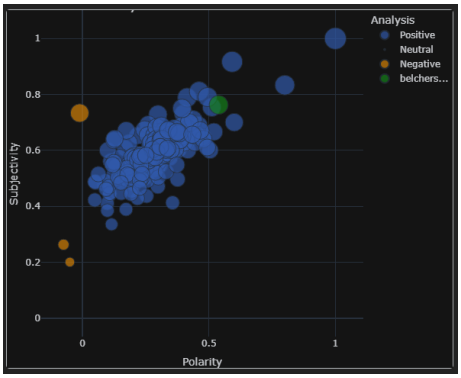


Fig. 11. Map showing all similar trails to the selected trail

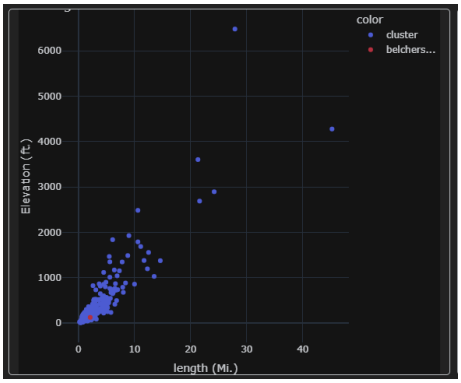
2) Visual Analysis of Sentiment for Selected Trail and Similar Trails: Several components were added to the Local view to give the user the ability to make an informed decision about selecting a trail for their global sentiment analysis. As can be seen in figure 12 (a), a scatter plot of all similar trails sentiment is shown, with the selected trail highlighted. Additionally shown in 12 (b), a scatter plot of the height versus length of all of the similar trails to that of the selected trail are visualized. Lastly, a word map can be seen in figure 12 (c) displaying the most common words in the written reviews of the selected trail. The aim of these three discussed plots is to give users the mechanism to make informed decisions on which trails to include in their global sentiment analysis.

3) Basic Information on Selected Trail: In addition to the information provided via Machine Learning *i.e.* Sentiment Analysis and Clustering, some basic information about the selected trail is provided.

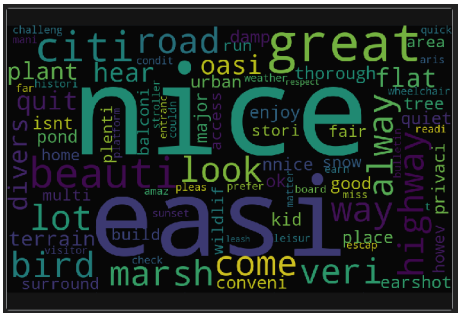
As shown in figure 13, information taken directly from the dataset is shown. This aims to give the user a more in depth/*real* feel about the trail there are looking at. The trail description contains an average of all of the star ratings associated with the given trail, a written description of the trail, and associated tags for that trail. The basic trail stats give the user a general idea of the trail itself with respect to height and elevation, as well as number of written reviews and star



(a) Local Sentiment Scatter Plot



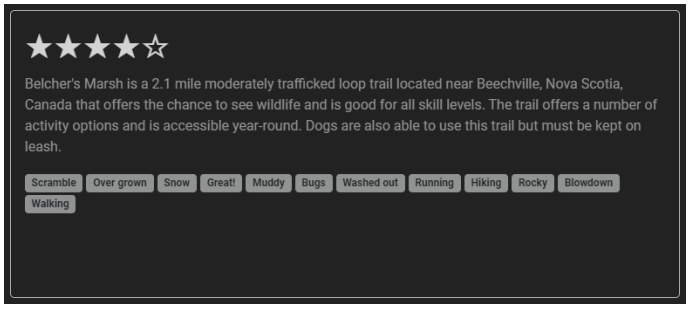
(b) Length vs. height for Similar trails



(c) Selected Trail Word Cloud

Fig. 12. Global Sentiment Analysis of selected trails

(a) Basic Trail Stats



(b) Trail Description

Fig. 13. Global Sentiment Analysis of selected trails

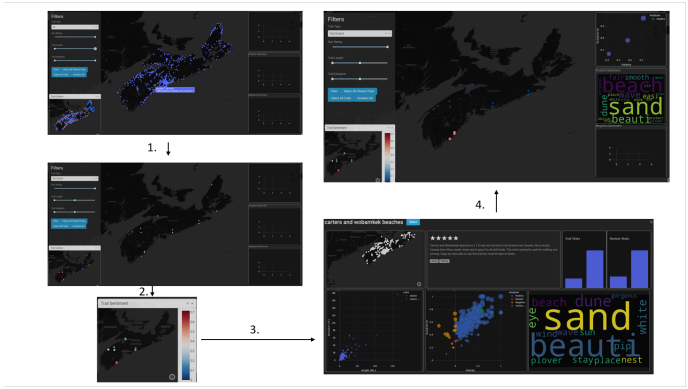


Fig. 14. The steps taken by Joe in his selection of trails for a weekend hiking trip

reviews. Users can hover over the bars to get an exact value for each of the above mentioned values. It is notable that the values displayed in the bar plots are normalized, this is to make the plots more meaningful, as without normalization height is always significantly larger than length; likewise with written reviews and star reviews.

V. Use Cases

To motivate the choice of visualizations, as well as demonstrate the workflow of the designed visual analytics tool, a sample use case will be provided.

Joe just moved to Nova Scotia and would like to plan a weekend hiking trip, as Joe is new to Nova Scotia he wants to make sure that he chooses the best trails for his first hiking trip. Additionally, Joe is new to hiking, so he would like to ensure that the trails he selects is not too difficult, i.e. not too much elevation or length. Joe also only prefers Out and Back trails because he likes taking everything in twice. Below his usage of the developed app is discussed and the process can be pictured in figure V.

1) First Joe opens up the developed app app and selects

filters that suits his needs - not too much length, nor elevation, a high rating and only consisting of out and back trails.
2) After filtering the results, a very refined set of trails are now visible, Joe notices there are there trails in the more South-Western part of Nova-Scotia with very positive sentiment associated with them.
3) Joe looks at those three trails individually, adding each of them one at a time, they all have high star reviews, they are in a good spot with respect to the similar trails around them for both length and elevation, and sentiment and some interesting popular words such as sand and beautiful are cropping up.
4) After selecting all three trails he gets a global view of the sentiment for those three trails, there are no negative reviews for all the trails, and the sentiment is relatively high for all three, one being very high, he expects this one to be his favourite.

In this case the developed app allowed Joe to speed up his process of choosing trails in a region close to him that all contained very positive reviews and met his needs in terms of difficulty. Sentiment Analysis and Clustering allowed Joe to speed up this process, as the alternative method of selection would be to try and read through a large number of written reviews, and try to piece similar trails together manually.

VI. Key Findings

Several key findings were made in the proces of developing this app. They can be listed as follows:

- Out of 712 trails, 595 were positively reviewed, while the remaining 117 were negatively reviewed. This helped us label the trails accordingly and map them on our Dashboard.
- Word Cloud portrays words like “beauty”, “great”, “nice”, etc. as the most prominent words in the positive reviews. Negative reviews encompass words such as “low”, “private”, “trespass”, etc. as the most frequent ones. This provided us a good understanding of the reviews for each trail.
- LDA Modelling on the filtered 15 negative review documents led to identifying a five-topic model with learning decay of 0.9, Log-Likelihood Score – 205.69020766995308 and perplexity metric - 634.7811 as the most optimal model.
- Topic models comprised of high probability words like “challeng”, “hard” and “wet”. These words can be traced back to their associated trails to understand which factors of the respective trails could be improved for better hiking experience.

VII. Conclusion and Future Directions

In conclusion a visual analytics tool was developed to streamline the process of interpretting the sentiment and clustering of hiking trails in Nova Scotia. Some interesting facts regarding the sentiment of the trails, and underlying clustering in the trails was discovered. In our future works we can extend our sentiment analysis by scraping the website for basic information of reviewers to visualize key observations of the type of reviews (positive and negative) against reviewer demographics. Furthermore, we can enhance our clustering by analyzing and categorizing features to determine more meaningful attributes. As an instance, there are three tags about dogs (i.e. dogs on leash, dog friendly, and no dog), or many tags describing the activity (e.g. walking, running, biking, hiking, etc.) Consequently, we can build a trail recommender system based on clustering trails applying the meaningful features.

References

[1] “13 Best Hikes in Canada: PlanetWare.” PlanetWare.com, https://www.planetware.com/canada/best-hikes-in-canada-cdn-1-221.htm.

[2] Costa, Anita. “The 10 Best Hikes in Canada.” Reader’s Digest Canada, Reader’s Digest Canada, 6 July 2021, https://www.readersdigest.ca/travel/canada/top-10-greatest-hikes-canada/.

[3] Horodyski, Kate. “The 12 Most INCREDIBLE Hikes to Take in Canada.” Culture Trip, The Culture Trip, 29 Dec. 2017, https://theculturetrip.com/north-america/canada/articles/the-12-most-incredible-hikes-to-take-in-canada/.

[4] Horodyski, Kate. “The 12 Most INCREDIBLE Hikes to Take in Canada.” Culture Trip, The Culture Trip, 29 Dec. 2017, https://theculturetrip.com/north-america/canada/articles/the-12-most-incredible-hikes-to-take-in-canada/.

[5] B. Liu, “Sentiment Analysis and Opinion Mining,” Accessed: Dec. 04, 2021. [Online]. Available: https://www.cs.uic.edu/ liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[6] S. Kunal, A. Saha, A. Varma, and V. Tiwari, “Textual Dissection of Live Twitter Reviews using Naive Bayes,” Procedia Computer Science, vol. 132, pp. 307–313, Jan. 2018, doi: 10.1016/j.procs.2018.05.182

[7] J. chavan, “NLP: Tokenization , Stemming , Lemmatization , Bag of Words ,TF-IDF , POS,” Medium, May 08, 2020. https://medium.com/@jeevanchavan143/nlp-tokenization-stemming-lemmatization-bag-of-words-tf-idf-pos-7650f83c60be (accessed Dec. 04, 2021).

[8] L. B. Shyamasundar and P. Jhansi Rani, “A Multiple-Layer Machine Learning Architecture for Improved Accuracy in Sentiment Analysis,” The Computer Journal, vol. 63, no. 3, pp. 395–409, Mar. 2020, doi: 10.1093/comjnl/bxz038

[9] A. Marco, “Stemming, Lemmatisation and POS-tagging with Python and NLTK,” Marco Bonzanini, Jan. 26, 2015. https://marcobonzanini.com/2015/01/26/stemming-lemmatisation-and-pos-tagging-with-python-and-nltk/ (accessed Dec. 04, 2021).

[10] D. A. Ostrowski, “Using latent dirichlet allocation for topic modelling in twitter,” in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Feb. 2015, pp. 493–497. doi: 10.1109/ICOSC.2015.7050858

[11] S.-W. Kim and J.-M. Gil, “Research paper classification systems based on TF-IDF and LDA schemes,” Hum. Cent. Comput. Inf. Sci., vol. 9, no. 1, p. 30, Aug. 2019, doi: 10.1186/s13673-019-0192-7.