Hashtag Relevancy in Twitter Posts

Problem Statement:

The challenge of hashtag relevancy is formulating a way to evaluate the suitability and efficacy of hashtags that are recommended for tweets on Twitter. Developing an algorithm to assess whether suggested hashtags are relevant to the tweet's content, context, and intent—as well as their ability to effectively engage the target audience—is the challenge at hand. To determine the most precise and effective method for hashtag recommendation, a framework for evaluating the relevance of hashtags produced by different recommendation engines in comparison to baseline approaches must also be developed.

Objective:

The objective is to develop a model for calculating the relevancy score of hashtags in a Twitter posts dataset using a pretrained BERT model and cosine similarity. This model will involve data collection and preprocessing, the utilization of a pretrained BERT model to create numerical embeddings for hashtags and tweet text, and the calculation of cosine similarity to determine relevancy. The goal is to assign relevancy scores to hashtags, with higher cosine similarity values indicating higher relevancy, and to evaluate the model's performance using appropriate metrics.

Literature Review:

[1] R. K. Recommending an Item from users Opinion or Sentiment on a given Hashtag using Twitter Data

This paper proposes an algorithm to recommend a product or service from users' opinions on a given hashtag using Twitter data. The paper discusses the importance of sentiment analysis in social media monitoring and how it can be used to learn about customer attitudes towards a product or service. The paper also highlights the use of hashtags in Twitter and how they can be used to categorize tweets and propagate trendy topics among millions of users. The paper proposes a novel algorithm for recommending an item from a given hashtag. The algorithm is called Hashtag Frequency-Inverse Hashtag Ubiquity (HF-IHU) and is a variation of the well-known TF-IDF that considers hashtag relevancy as well as data sparseness, which is one of the key challenges in analyzing microblog data. The algorithm relies on two central data structures that are compiled from a large number of tweets: the first is a term to hashtag-frequency-map (THFM), and the second is the converse—a hashtag frequency-map (HFM). The paper also discusses the challenges of processing large amounts of Twitter data and how the proposed algorithm can be used to provide personalized recommendations for a user. The paper concludes by stating that the proposed approach is more robust towards unstructured datasets than other methods, since pairwise similarities are exploited and words are intercorrelated based on the HF-IHU algorithm.

Hashtag relevancy is an important factor in the proposed algorithm, as it considers the frequency with which a hashtag appears with a given term (the hashtag frequency) and the hashtag ubiquity, which discounts hashtags that are prevalent in all contexts and rewards hashtags that are tightly associated with a narrow subset of terms. The algorithm scores hashtags in the data set to find personalized recommendations for a user based on these factors.

[2] Hashtag recommendation approach based on content and user characteristics

The problem of hashtag suggestion on Twitter is addressed in this study. In order to provide users with relevant hashtag recommendations based on their changing interests and the content of real-time tweets, the authors provide a novel approach that combines online Twitter-User LDA and incremental biterm topic model (IBTM).

The study identifies the shortcomings of the current hashtag recommendation techniques and divides them into two categories: those that take user interests into account and those that depend on tweet similarity. They stress that the efficacy of earlier approaches has been hampered by their frequent use of static topic models that are inappropriate for streaming tweets and disregard user interests.

The authors highlight two major issues with recommendation systems: the dearth of hashtags and the significance of hashtag relevancy. Additionally, they recognise that hashtag tagging is subjective and that a tweet's main ideas may not always be captured by a single hashtag.

The authors suggest the User-IBTM technique, which combines online Twitter-User LDA with IBTM, as a solution to these problems. Online Twitter-User LDA records the dynamic interests of users, while IBTM investigates the subject distribution of tweet material in real time. The objective of the User-IBTM technique is to offer more specific and accurate hashtag recommendations by taking into account both user interests and tweet content.

In summary, the paper contributes to the field of Twitter hashtag recommendation by introducing an innovative method that considers both user interests and live tweet content. Experimental results on real-world Twitter data demonstrate the superiority of their approach over several baseline methods in terms of hashtag recommendation accuracy.

[3] Suggest what to tag: Recommending more precise hashtags based on users' dynamic interests and streaming tweet content

The paper "Suggest What to Tag: Recommending More Precise Hashtags Based on Users' Dynamic Interests and Streaming Tweet Content" proposes a method for automatic hashtag recommendation in Twitter that considers both dynamic user interests and streaming tweet content. The authors use topic models to discover the latent topics of each single tweet and then select suitable keywords for hashtag recommendation. The paper highlights the challenges of hashtag recommendation, including the scarcity of hashtags and the dynamic nature of hashtags. The authors also discuss the characteristics of hashtags, such as their length, frequency, quantity in one tweet, and diversity. The paper provides a literature review of previous studies on hashtag recommendation, including methods based on traditional recommender systems, topic models, and supervised topic models. The authors argue that their proposed method outperforms several other baseline methods and can suggest more precise hashtags. The paper is informative and provides a comprehensive overview of the topic of hashtag relevancy.

[4] Design and evaluation of a twitter hashtag recommendation system

This paper proposes an automatic hashtag recommendation system that helps users find new hashtags related to their interests. The proposed system uses a variation of the well-known TF-IDF approach called Hashtag Frequency-Inverse Hashtag Ubiquity (HF-IHU) to score hashtag relevancy while also taking into account data sparseness of Twitter data set. The system was evaluated on a large Twitter data set and demonstrated that it performed better than other methods that rely only on hashtag popularity. The proposed method successfully yields relevant hashtags for user's interest and that recommendations are more stable and reliable than ranking tags based on tweet content similarity. The study compares three different hashtag ranking methods and shows that the proposed method outperforms the other two methods. The study also observes the challenge of ranking hashtags based on tweet similarity and recommends hashtags existing in similar tweets due to the sparseness of hashtags. The proposed system attempts to retrieve relevant and emerging hashtags in the data set, while other approaches limit the suggestions to general topics. The study concludes that the proposed system is effective in recommending personalized trending hashtags based on users' tweets. However, the study does not provide a detailed analysis of hashtag relevancy.

[5] Hashtag recommendation for short social media texts using word-embeddings and external knowledge

This paper discusses the problem of hashtag relevancy in short social media texts, particularly tweets. The authors introduce a five-step approach for hashtag recommendation, which includes extracting extrinsic features from external sources like Wikipedia, selecting and processing features, generating candidate hashtags, computing user influence scores, and recommending candidate hashtags.

The paper extends previous research in hashtag recommendation by referencing studies that explored different methods such as topic models, graph-based ranking, and collaborative filtering. The authors emphasize the importance of considering both internal and external semantics to cluster short texts effectively.

This paper offers a comprehensive solution to improve hashtag recommendations for short social media texts, addressing relevancy by integrating extrinsic features from external sources. Their system combines various techniques to generate candidate hashtags and calculate user influence scores, resulting in more contextually relevant hashtag suggestions for tweets.

[6] A hashtag recommendation system for twitter data streams

This paper focuses on the problem of recommending personalized trending hashtags on Twitter. The authors introduce a novel approach that addresses the challenge of hashtag relevancy and data sparseness in microblog data analysis.

They present the Hashtag Frequency-Inverse Hashtag Ubiquity (HF-IHU) ranking scheme, which combines hashtag relevance and popularity. This approach proves to be more stable and reliable than content similarity-based methods in recommending relevant hashtags for users.

The paper highlights the limitations of existing hashtag recommendation systems, such as recommending predefined hashtags or popular topics, which may not be suitable for users seeking emerging and personalized hashtags. The authors evaluate their system using a TREC

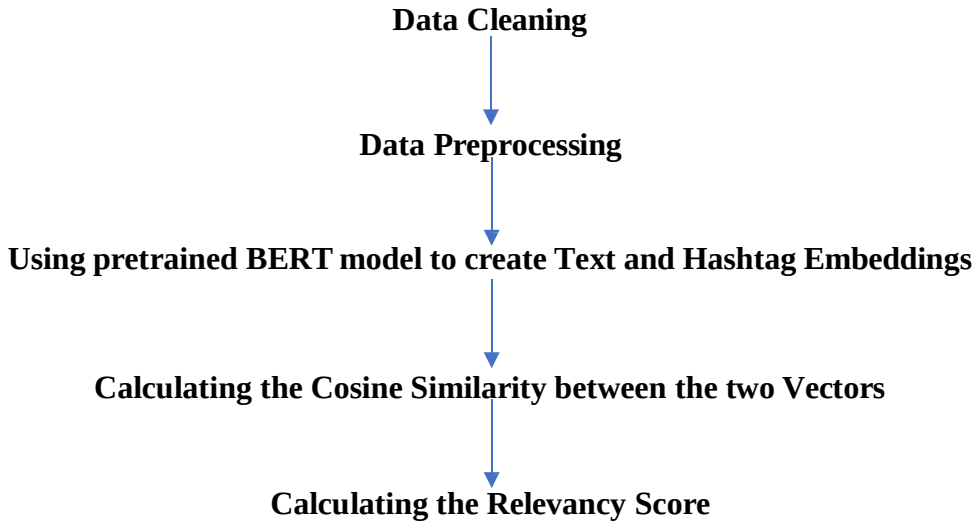
2011 Microblog Track dataset and show that it outperforms other recommendation methods in terms of hashtag recall.

This paper contributes to the field of personalized hashtag recommendation on Twitter by addressing existing system limitations. Their innovative approach, HF-IHU, effectively considers both relevance and popularity, resulting in more accurate and personalized hashtag recommendations for users' interests.

Dataset Description:

The dataset used is downloaded from Data World. Out of the different columns present the 'text' column is the one which is important as this is the column which contains the tweet post including the hashtags. The data in this column is used for analysis and pre-processing for cleaning and cleaned dataset is prepared. The cleaned dataset is further used for finding the relevancy score for the hashtags used.

Flowchart:



Data Preprocessing and Predictive Modelling:

Data Cleaning:

1. The original dataset is loaded into Jupyter Notebook.
2. Irrelevant columns are dropped and only the 'text' column is kept.
3. Null values are removed.
4. URLs are removed as they are not relevant while considering the hashtags.
5. Removing the retweets mentioned, the tweet post is considered rather than considering whose post was retweeted.
6. Extracting the hashtags in a different column and creating a column which contains only the tweet text without any hashtags.
7. Creating columns for extracting the accounts mentioned in the tweet and for tweet post without any account mentions.
8. Detecting the language of the post and creating a column which gives 'en' for English language posts and 'unknown' for other language posts.

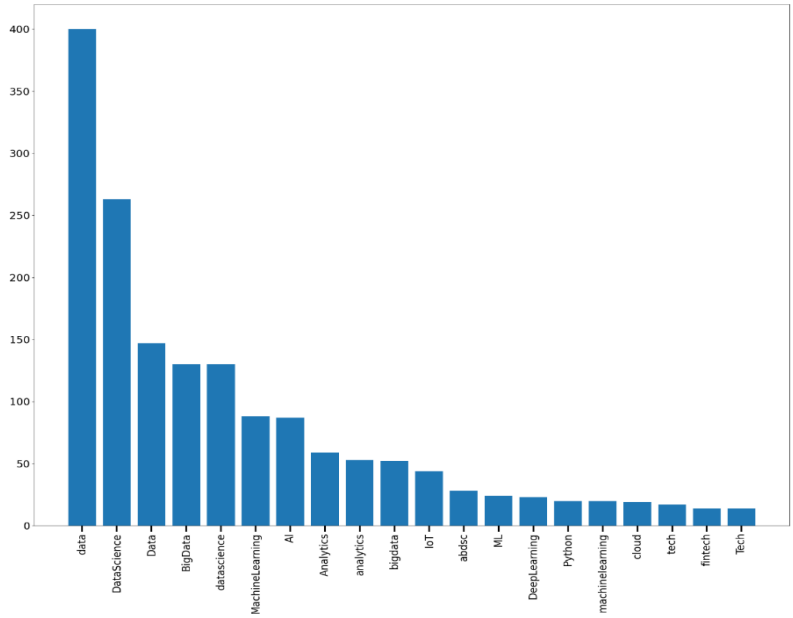
9. Creating a cleaned dataset which is the subset of the above dataset and containing only the English language tweet post.

Data Preprocessing:

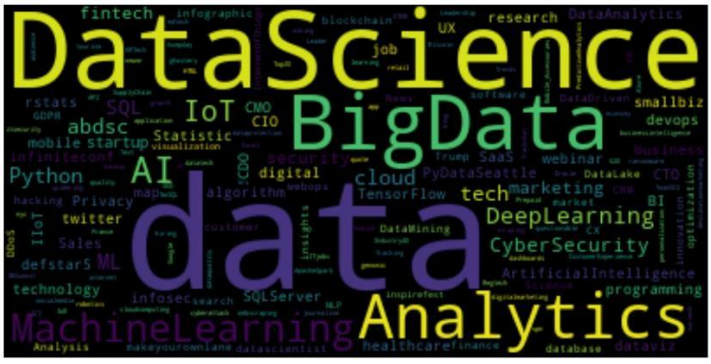
1. Handling lines with unexpected number of fields.
2. Removing duplicates and null values.
3. Cleaning the tweet text and hashtags.
4. Tokenizing the text and hashtags using BERT tokenizer from the Hugging Face Transformers library.
5. Combining the text and hashtags tokens.

Data Visualizations:

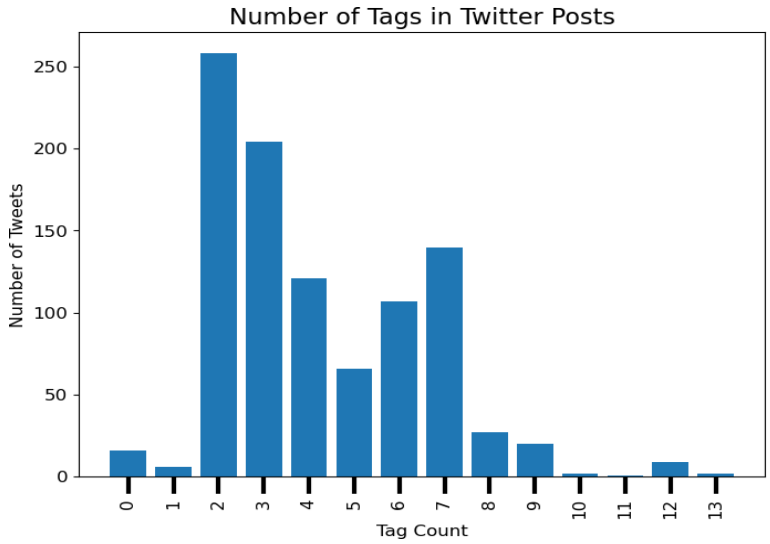
1. Word count of different hashtags used in the whole dataset.



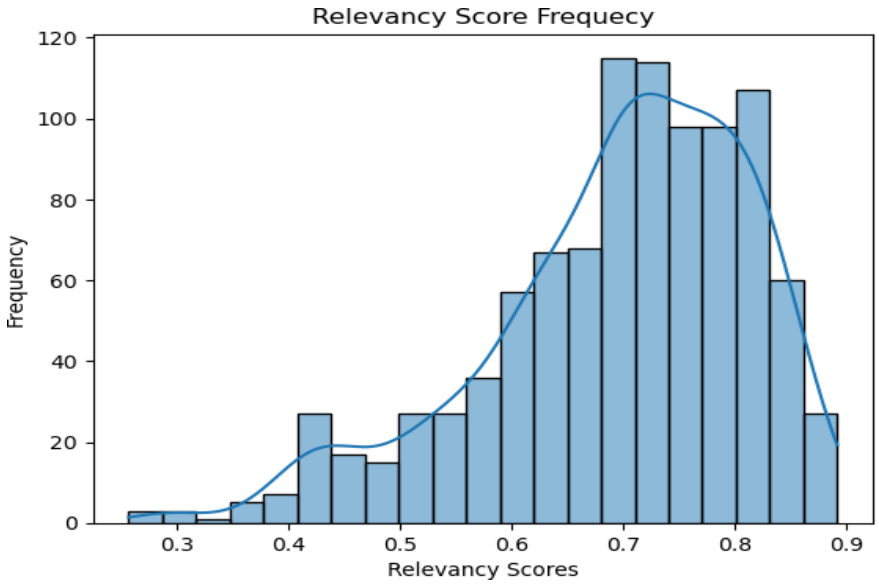
2. Word Cloud for the different hashtags used.



3. Number of tags used in different posts.



4. Relevancy Scores.



Results and Observations:

1. The pre-processing and tokenization techniques helped in developing a model that gives the Relevancy Score for the hashtags used in each tweet post.
2. Tokenization was performed using the BERT tokenizer from the Hugging Face Transformers library.
3. This project used the Pretrained BERT model for getting the word embeddings.
4. Model used Cosine similarity to calculate the similarity between the vectors and the relevancy score of the hashtags.
5. The relevancy score for each post is calculated.
6. Most used hashtags: #data, #DataScience, #Data, and #BigData. We can observe this from the word cloud generated.
7. Most of the hashtags had relevancy score between 0.70 and 0.80.

Conclusion and Recommendations:

Hashtag relevancy in Twitter posts plays a pivotal role in engaging audiences and maximizing social impact without incurring additional costs. In this project, we employed a multifaceted approach, including tokenizing the text and hashtags using the BERT tokenizer from the Hugging Face Transformers library, generating word embeddings using a pretrained BERT model, calculating cosine similarities, and ultimately computing the relevancy scores. The hashtags were considered to be relevant or not based on a threshold relevancy score of 0.50. If the score is less than 0.50 then the tags are considered to be non-relevant, else they were considered to be relevant.

Recommendations:

1. *Cross-Platform Integration*: The success of hashtag recommendation systems is not limited to Twitter. The lessons learned here can be applied to other social media platforms like Facebook, Instagram, LinkedIn, Pinterest, and TikTok. Developing hashtag recommendation systems tailored to these platforms could amplify user engagement and content visibility.
2. *Advanced Pre-processing Techniques*: Exploring advanced pre-processing techniques, such as feature engineering, data reduction, feature selection, and feature scaling, can further enhance the accuracy and performance of hashtag relevancy models. These techniques can help in fine-tuning the model's ability to understand and recommend hashtags effectively.
3. *Sentiment Analysis*: For an even more comprehensive approach, incorporating sentiment analysis into hashtag relevancy assessment can be highly beneficial. Visualizing the sentiment associated with different hashtags through tools like bar charts allows content creators to align their messaging with the prevailing sentiment, thereby improving audience engagement and resonance.

In conclusion, hashtag relevancy remains a critical aspect of social media strategy. Leveraging advanced techniques and extending the lessons learned to other platforms, along with considering sentiment analysis, can propel hashtag relevancy models to new heights, enhancing their value in the ever-evolving landscape of social media communication.

Code link: Repo Link

References:

[1] Indira, M. D., & Kumar, R. K. Recommending an Item from users Opinion or Sentiment on a given Hashtag using Twitter Data.

[2] Tran, V. C., Hwang, D., & Nguyen, N. T. (2018). Hashtag recommendation approach based on content and user characteristics. Cybernetics and Systems, 49(5-6), 368-383.

[3] Li, J., & Xu, H. (2016). Suggest what to tag: Recommending more precise hashtags based on users' dynamic interests and streaming tweet content. Knowledge-based systems, 106, 196-205.

[4] Otsuka, E., Wallace, S. A., & Chiu, D. (2014, July). Design and evaluation of a twitter hashtag recommendation system. In Proceedings of the 18th International Database Engineering & Applications Symposium (pp. 330-333).

[5] Kumar, N., Baskaran, E., Konjengbam, A., & Singh, M. (2021). Hashtag recommendation for short social media texts using word-embeddings and external knowledge. Knowledge and Information Systems, 63, 175-198.

[6] Otsuka, E., Wallace, S. A., & Chiu, D. (2016). A hashtag recommendation system for twitter data streams. Computational social networks, 3, 1-26.