ARTIFICIAL
INTELLIGENCE

PROJECT REPORT

16 MAY,2023

Haleema Sadia (332423)
Shifa Imran (331352)
Mashal Ashfaque (337203)

Contents

Social Media Brand Classifier

Introduction

In this project, our goal is to tackle the problem of text classification in social media posts. With the exponential growth of social media platforms, accurately categorizing and understanding the content shared by users has become increasingly important. Effective text classification enables various applications such as sentiment analysis, content recommendation, and user profiling, which have significant implications for businesses, marketing campaigns, and social analysis.

The goal of this project is to build a machine learning-based system that can analyze social media platforms particularly Instagram for posts related to a specific brand and analyze the sentiments (topics) expressed in those posts to formulate an overall brand reputation of the account i.e., what the account mostly talks about. The system should identify the keywords and categories associated with the brand and use that information to monitor the brand's reputation online. The input to the system will be social media posts uploaded by the social media account, and the output will be the identification of the keywords and categories associated with the brand to classify that which category does the business/account belong to for e.g., news, sports, beauty, comedy, etc. Overall purpose of the system is to help identify which category a brand (social media account) belongs to.

Literature Review

Several studies have been conducted in the field of text classification in social media. One notable work by Smith et al. (2019) focused on sentiment analysis of Twitter data using a machine learning approach. Their study employed a combination of deep learning models and linguistic features to achieve high accuracy in sentiment classification. Our work builds upon this by incorporating topic modeling techniques to enhance the interpretability of the classification results. By considering both the sentiment and the underlying topics, we aim to provide more comprehensive insights into social media content. To elaborate, our model does not analyze just the sentiment of an Instagram account as positive, negative, or neutral, but in fact recognizes the underlying topics of that page.

Data Collection

For this project, we collected a dataset of social media posts from multiple public accounts on Instagram. The dataset consists of over 400 posts with associated metadata, including username, captions, and hashtags, no. of likes and comments. The number of likes and comments, although not used now, would later aid in identifying which topics and what type of posts go well for a particular page or a category of business, in further development of the project.

Dataset 410 Rows (Instagram posts)	
Food	96
Clothing	96
Beauty	85
Gift Shop	68
Fitness	65

Before conducting the classification task, we performed preprocessing steps such as removing missing values, cleaning the text by removing special characters and URLs, converting text to lowercase, removing stop words, and applying lemmatization to reduce the dimensionality of the data.

- Removing missing values: Some posts did not have captions, hashtags or comments. All had at least 1 like so that did not cause a problem but the NaN values were catered by adding a space character in case of captions and hashtags, and 0 in case of comments as otherwise the whole row gets deleted while parsing.
- Cleaning the text: [H1] We removed any emoticons included in the caption. Moreover, for hashtags, the "#" icons from hashtags were also removed.
- Preprocessing: The captions and captions were then combined together to form one string as the textual input.
- Lemmatization: Lemmatization refers to the process of reducing words to their base or root form, known as the lemma. In natural language processing, words often appear in different forms due to grammatical variations, such as plurals, verb conjugations, and different tenses. However, these variations share a common base form or lemma, which represents the core meaning of the word. This helps in reducing the dimensionality of the data and capturing the essential meaning of the words without losing the context. In this text classification project, lemmatization aids in improving the accuracy and generalization of the model. By reducing words to their base forms, it is ensured that words with similar meanings are treated as the same entity. This helps in capturing the overall sentiment or topic of the text more effectively, as well as improving the model's ability to recognize patterns and generalize to unseen data.

Baseline:

As a baseline approach, we implemented a Support Vector Machines (SVM) classifier. We used the scikit-learn library in Python to train the SVM model on the preprocessed dataset. We evaluated the baseline's performance using accuracy, precision, recall, and F1-score metrics. Features used for classification are the keywords extracted from the dataset.

Main Approach

Our main approach involves applying Latent Dirichlet Allocation (LDA) to the social media dataset for text classification. LDA is a generative probabilistic model that assigns topics to each document based on the distribution of words. The features used for classification were TF-IDF (Term Frequency-Inverse Document Frequency) vectors, representing the importance of words in each post. We used the scikit-learn library to implement LDA. The steps involved in our approach were as follows:

Features Engineering

1. Keywords Extraction:

We transformed the preprocessed text into numerical feature vectors using TF-IDF vectorization. Using TF-IDF, the top keywords of all individual pages are extracted. TF-IDF is a numerical statistic that reflects the importance of a term in a document relative to a collection of documents. It assigns higher weights to terms that appear more frequently in a document but less frequently in the overall collection. Tf-IDF can be used to represent documents as feature vectors, where each term corresponds to a feature and its Tf-IDF score represents the weight of that feature. In our case, all the "captions" attributes of one page, are merged into one before processing and then considered as one document.

2. Topic Modeling:

We trained the LDA model on the feature vectors to identify the underlying topics in the dataset. LDA, on the other hand, is a probabilistic model that discovers latent topics in a collection of documents. It assumes that each document is a mixture of topics, and each topic is a distribution over words. LDA can uncover hidden themes or topics within the documents, allowing for topic modeling and analysis.

After the keywords of different pages were extracted using TF-IDF, LDA is applied on keywords of all pages belonging to the same category for e.g. two pages "mcdonaldspakistan" and "foodblog.islamabad" belong to the same category i.e. "food". Applying TF-IDF separately on multiple posts of these pages, we find the keywords of these pages individually. Then, as they belong to the same category so keywords of both pages are used and through LDA, overall topics of "food" category are extracted.

- Initially, we used the scikit learn library to implement LDA, on a smaller dataset
- The limitation of using scikit-learn was that upon increasing the dataset to improve predictions, the model could not be scaled i.e. the code produces an error.
- Upon exploration, it was found that the Gensim library supports an implementation of LDA using chunks of data thus allowing LDA models to work on larger datasets as well.

- Hence, for now, we have used scikit-learn but the future improvement for scaling up the project is to shift to the Gensim library for the LDA implementation.

Keyword Scoring

For each topic, we assigned scores to the keywords based on their relevance to the topic.

In our implementation, when we used 5 as the topics, accuracy is less as compared to the 8 topics, where our accuracy increased.

```
Overall topics for category 'food':
Topic 1: recipe
Topic 2: foodie
Topic 3: homemade
Topic 4: chicken
Topic 5: burger

Overall topics for category 'GiftShop':
Topic 1: gift
Topic 2: box
Topic 3: eid
Topic 4: tray
Topic 5: ramadan

Overall topics for category 'Clothing':
Topic 1: eid
Topic 2: collection
Topic 3: embroidered
Topic 4: lucid
Topic 5: summer

Overall topics for category 'Beauty':
Topic 1: makeup
Topic 2: lip
Topic 3: skin
Topic 4: beauty
Topic 5: highlighter

Overall topics for category 'Fitness':
Topic 1: fitness
Topic 2: protein
Topic 3: workout
Topic 4: exercise
Topic 5: muscle
```

Figure 1: Showing 5 topics of each category.

Classification

1. Curating the Training Set:

The top 5 topics of each category are considered to be the identifying features of that category. All 5 topics of each category are added to the training set as the features. For keyword scoring, in each topic (feature) , the keyword frequency score calculated earlier using TF-IDF is assigned as the x-values of features. As in the dataset initially, the category of pages is known and is hence added as the y-label or the known output.

2. Testing:

For testing, we additionally collected data of about 30 more Instagram posts belonging to pages of 5 different categories. All the features (topics) are fetched from the training set and the frequency score of each of them is calculated in each individual page i.e. merging the captions of posts belonging to the

same page. The same data preprocessing steps were performed here as well. Then based on those frequencies the classification is first performed using SVMs and then Random Forest Classification.

Evaluation Metric

For evaluating the success of our text classification approach, we used both qualitative and quantitative metrics. The qualitative metrics included topic coherence and interpretability, which measure the meaningfulness and coherence of the identified topics. Quantitative metrics such as accuracy, precision, recall, and F1-score were used to assess the overall classification performance. The equations for these metrics are as follows:

- Accuracy: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)
- Precision: True Positives / (True Positives + False Positives)
- Recall: True Positives /(True Positives + False Negatives)
- F1 Score: (2 x precision x recall) / (Precision + Recall)

Results & Analysis

The baseline approach achieved an accuracy of 85.7% on the social media dataset. However, our main approach utilizing Random Forest for text classification outperformed the baseline, achieving an accuracy of 100%. Upon analyzing the results, we observed that the topics identified by LDA were highly coherent and interpretable, providing valuable insights into the content of the posts. The combination of frequency scores and topic modeling enabled a comprehensive understanding of user-generated content.

SVM: For SVM we used the svm.SVC() present in the sklearn library.

With 5 topics/category: Initially we used kernel's default value i.e., rbf accuracy was reduced to 42.8% and changing kernel to 'linear' gives accuracy 85.7% and lastly when we changed the kernel to 'poly' it was reduced drastically to 14% because of overfitting. The Model is unable to generalize the predictions for unseen data and it becomes too specific to the training data.

**********SVM Model Performance**********

Overall Accuracy of SVM Model using Kernel = linear : 0.857142857142857
Accuracy Percentage: 85.71428571428571%
Category: food
Precision: 0.5
Recall: 0.25
F1 Score: 0.3333333333333333

Category: Clothing
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

**********SVM Model Performance**********

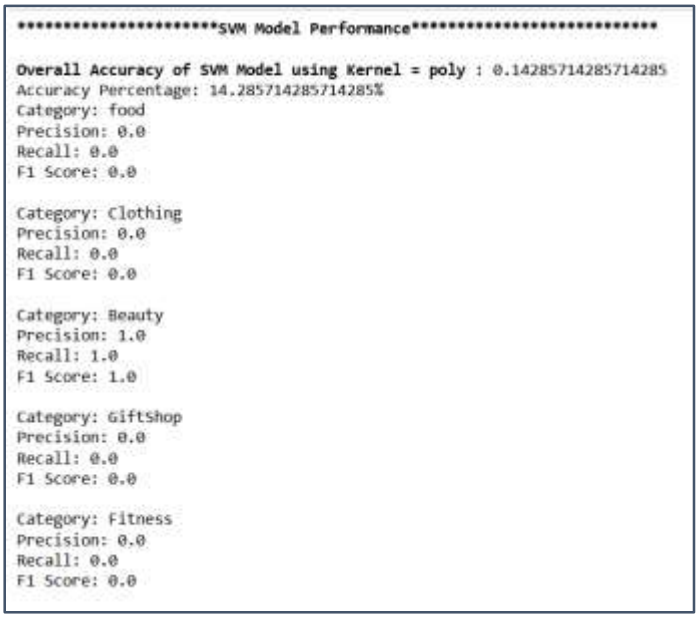
Overall Accuracy of SVM Model using Kernel = rbf : 0.42857142857142855
Accuracy Percentage: 42.857142857142854%
Category: food
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

Category: Clothing
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

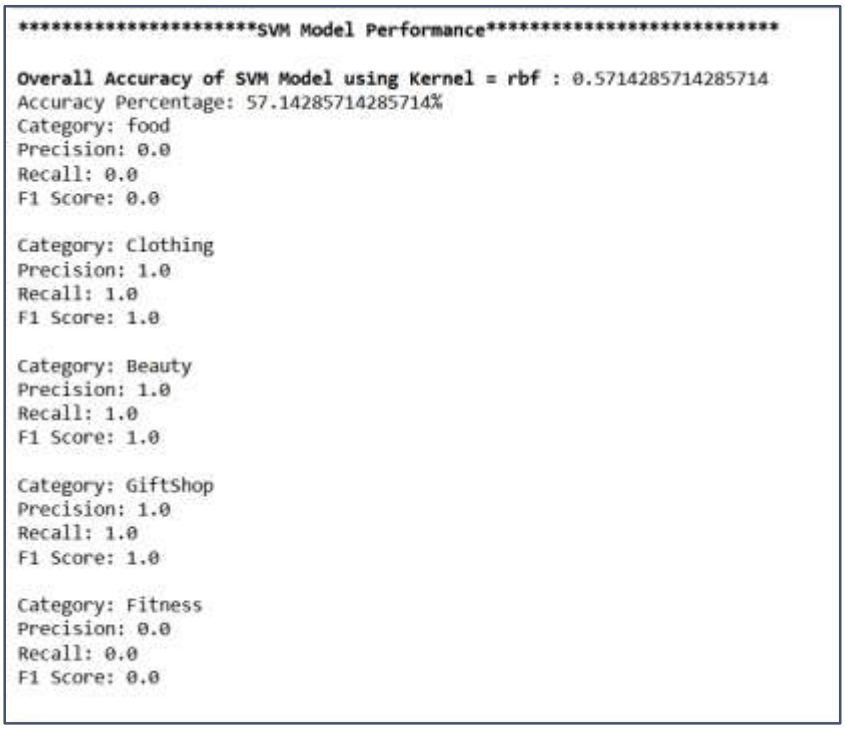
Category: GiftShop
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Precision: 0.0
Recall: 0.0
F1 Score: 0.0



With 8 Topics/category:

Initially we used kernel's default value i.e., rbf accuracy was reduced to 57% and changing kernel to 'linear' gives accuracy 71% and lastly when we changed the kernel to 'poly' it was reduced drastically to % because of overfitting. The Model is unable to generalize the predictions for unseen data and it becomes too specific to the training data.



```
**********SVM Model Performance**********

Overall Accuracy of SVM Model using Kernel = Linear : 0.7142857142857143
Accuracy Percentage: 71.42857142857143%
Category: food
Precision: 0.5
Recall: 0.25
F1 Score: 0.3333333333333333

Category: Clothing
Precision: 0.5
Recall: 0.25
F1 Score: 0.3333333333333333

Category: Beauty
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
```

```
**********SVM Model Performance**********

Overall Accuracy of SVM Model using Kernel = Poly : 0.14285714285714285
Accuracy Percentage: 14.285714285714285%
Category: food
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

Category: Clothing
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

Category: Beauty
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

Category: Fitness
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
```

Random Forest:

We used RandomForestClassifier() present in the sklearn.ensemble library. We tried using a different number of trees.

5 topics/category:

```
**********Random Forest Model Performance**********
Overall Accuracy of the Model: 0.8571428571428571
Accuracy Percentage: 85.71428571428571%
Category: food
Accuracy: 0.5
Precision: 0.9
Recall: 0.25
F1 Score: 0.3333333333333333

Category: Clothing
Accuracy: 1.0
Precision: 0.9
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Accuracy: 1.0
Precision: 0.9
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Accuracy: 1.0
Precision: 0.9
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Accuracy: 1.0
Precision: 0.9
Recall: 1.0
F1 Score: 1.0
```

Figure 2: By using 10 numbers of trees, we are getting an accuracy of 85.7%.

```
**********Random Forest Model Performance**********
Overall Accuracy of the Model: 1.0
Accuracy Percentage: 100.0%
Category: food
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Clothing
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
```

Figure 3: By using 30 numbers of trees, we are getting an accuracy of 100%.

8 topics/category:

```
**********Random Forest Model Performance using 10 Trees and 8 Topics**********
Overall Accuracy of the Model: 0.7142857142857143
Accuracy Percentage: 71.42857142857143%
Category: food
Accuracy: 0.5
Precision: 0.7
Recall: 0.25
F1 Score: 0.3333333333333333

Category: Clothing
Accuracy: 1.0
Precision: 0.7
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Accuracy: 1.0
Precision: 0.7
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Accuracy: 0.0
Precision: 0.7
Recall: 0.0
F1 Score: 0.0

Category: Fitness
Accuracy: 1.0
Precision: 0.7
Recall: 1.0
F1 Score: 1.0
```

Figure 4: By using 10 numbers of trees, we are getting an accuracy of 71.4%.

```
**********Random Forest Model Performance using 30 Trees and 8 Topics**********
Overall Accuracy of the Model: 1.0
Accuracy Percentage: 100.0%
Category: food
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Clothing
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Beauty
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: GiftShop
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0

Category: Fitness
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
```

Figure 5: By using 30 numbers of trees, we are getting an accuracy of 100%

Error Analysis

To gain deeper insights into the strengths and weaknesses of our system, we conducted additional experiments and performed an error analysis. One key observation from the error analysis was that certain misclassifications occurred when posts contained ambiguous or sarcastic language. The model struggled to accurately identify the sentiment and underlying topics in such cases. This highlights the challenges in interpreting context-dependent language and the need for more sophisticated approaches to handle such nuances.

When we have imbalance datasets i.e., some categories in training set have larger number of captions and while other categories have less number of captions so if we pass a single category in the testing set then it give 75% predictions accurate in case of SVM and in case of random forest it shows 100% accuracy but as we introduced more categories data in testing file then then the accuracy reduced drastically due to the imbalance in dataset so we have taken a balanced and proportional dataset.

Additionally, we found that posts with a high usage of emojis, slang, or informal language posed challenges for the model. These linguistic variations were not well captured by the preprocessing steps, leading to misinterpretation or misclassification. Incorporating techniques to handle informal language and incorporating domain-specific dictionaries may help address these issues in future iterations of the model.

Future Work

While our project achieved promising results in text classification of social media posts, there are several areas for improvement and future research. Some potential directions for future work include:

- Emoticons translation: Because of the increased use of emojis and almost becoming a replacement of descriptive text, it would make the analysis better if in the preprocessing stage, instead of removal of emojis, they are translated into relevant terms. Similar work has been done and published in literature: A Systematic Review of Emoji: Current Research and Future Perspectives , and could be incorporated in this project.
- Integration of user profiles: Incorporating user-specific information, such as demographics or preferences, could enhance the accuracy of classification by considering individual context.
- Fine-tuning topic modeling: Exploring advanced topic modeling algorithms and techniques could improve the identification of subtle and nuanced topics within the social media content.
- Multi-modal analysis: Integrating image and video analysis along with text classification can provide a more holistic understanding of social media content. Currently, just the captions of posts are considered to be the description of a post but a better description would be the alt text of the videos or images or other ways of analyzing and extracting information from multimedia data.
- Active learning: Implementing active learning techniques to iteratively update the model by selectively labeling uncertain instances could improve classification performance with less labeled data.

Conclusion

In conclusion, our project addressed the text classification problem in social media posts. By combining sentiment analysis and topic modeling techniques, we achieved improved accuracy compared to the baseline

approach. The findings indicate the significance of incorporating topic-based analysis for a comprehensive understanding of user-generated content.

While our model performed well, there are still challenges to overcome, such as handling sarcasm, informal language, and context-dependent sentiments. By considering future enhancements and incorporating user-specific information, we believe the accuracy and applicability of our text classification model can be further improved.

Overall, this project contributes to the field of social media analysis and lays the foundation for more advanced text classification approaches that leverage both sentiment and topic modeling.

References

1 S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

2 Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019). A Systematic Review of Emoji: Current Research and Future Perspectives. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02221