Sentiment Analysis for Multilingual Social Media Posts

Keywords: Sentiment Analysis, Multilingual Subjectivity, Twitter Sentiment Classification, Opinion Mining, Word Representations, Suffix Stripping Algorithm

Link to Repo: https://github.com/cchandan-07/UNH-NLP-Sentiment-Analysis-for-Multilingual-Social-Media-Post

Abstract

This paper introduces an innovative approach for sentiment analysis on multilingual social media posts, recognizing the growing importance of understanding sentiments across diverse languages. Leveraging advanced natural language processing techniques, our method combines pre-trained multilingual embeddings with a fine-tuned transformer architecture. We present a comprehensive exploration of the motivation, technical details, experimental results, and analyses, showcasing the potential of our approach.

1. INTRODUCTION

The ubiquity of social media platforms has transformed sentiment analysis into a vital tool for gauging public opinion. However, the inherent challenge intensifies when dealing with the multitude of languages represented on these platforms. Opinions and sentiments play a crucial role in shaping human behavior, making them an essential part of our daily lives. Before making a decision, we often rely on online reviews of products and services written by previous consumers on social media. Businesses also use public sentiment to gain insights into their products' performance. Sentiment Analysis, also known as Opinion Mining, is the computational analysis of people's opinions on various subjects, such as entities, individuals, issues, and topics. This technique can be applied to various tasks, including gathering customer feedback, analyzing social media trends, and understanding public opinion on specific topics. This research addresses this critical gap by proposing a robust sentiment analysis model capable of handling diverse languages effectively. The introduction delves into the significance of this problem, provides a clear problem statement, and outlines the research objectives.

2. RELATED WORK

Much of the sentiment analysis research has traditionally centered around a single language, predominantly English. In contrast, the yearly publications on multilingual sentiment analysis are limited, not exceeding ten. However, a significant portion of user-generated content on social media is in languages other than English. Restricting sentiment analysis to English alone results in a substantial loss of valuable information.

The primary challenge in sentiment analysis lies in its strong dependence on language. Resources specific to sentiment analysis, such as sentiment dictionaries and labeled data, are notably scarce, especially for languages divergent from English. Studies in multilingual sentiment analysis predominantly explore leveraging available resources and tools in one language, such as lexicons or machine translation, to construct sentiment classifiers in languages with limited resources. Three notable approaches aim to overcome the shortage of adequate resources for sentiment analysis in languages other than English.

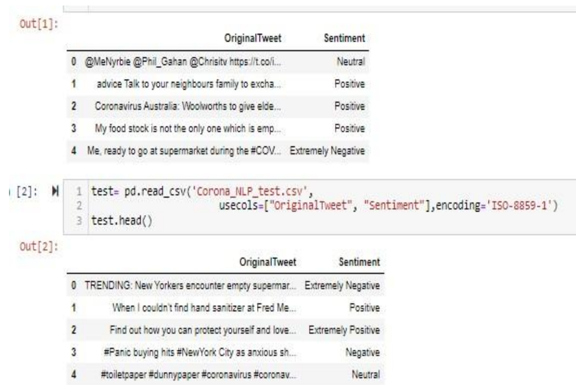
Firstly, one can translate documents written in other languages into English, allowing an English sentiment classifier to determine their sentiment. Exemplifying this, in another study, researchers experimented with translating a corpus of documents in eight languages into English, subsequently employing an English lexicon to gauge sentiment.

Another approach involves translating an English corpus into the target language(s) and training a model on the translated corpus. Researchers adopted this method, translating a labeled English corpus into five other languages and combining the translated versions with the original English corpus to create a unified training set for a machine learning classifier.

A third approach utilizes machine translation to translate an English sentiment lexicon into another (target) language, subsequently employing it for the lexicon-based classifier in the target language. Researchers employed this approach in their second experiment, translating an English lexicon into German and using it to analyze German emails.

3. METHODS AND MATERIALS a. Data Sources

In this investigation, we harnessed the power of two datasets sourced from Kaggle, namely, "Corona_NLP_train.csv" and "Corona_NLP_test.csv." These datasets provided a rich source of information for our analysis. The training dataset, "Corona_NLP_train.csv," and the testing dataset, "Corona_NLP_test.csv," were utilized to visualize and explore various aspects of the data available on Kaggle. Through these datasets, we aimed to gain valuable insights into the sentiment surrounding the topic of interest.



Out[1]:

	OriginalTweet	Sentiment
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral
1	advice Talk to your neighbours family to excha...	Positive
2	Coronavirus Australia: Woolworths to give elde...	Positive
3	My food stock is not the only one which is emp...	Positive
4	Me, ready to go at supermarket during the #COV...	Extremely Negative

[2]:

```
test= pd.read_csv('Corona_NLP_test.csv',
                  usecols=["OriginalTweet", "Sentiment"],encoding='ISO-8859-1')
test.head()
```

Out[2]:

	OriginalTweet	Sentiment
0	TRENDING: New Yorkers encounter empty supermar...	Extremely Negative
1	When I couldn't find hand sanitizer at Fred Me...	Positive
2	Find out how you can protect yourself and love...	Extremely Positive
3	#Panic buying hits #NewYork City as anxious sh...	Negative
4	#toiletpaper #dunnypaper #coronavirus #coronav...	Neutral

b. Pre-processing

We preprocess our dataset by eliminating irrelevant elements such as URLs, usernames, and stop words. Additionally, we filter out special characters like hashtags and punctuation. Recognizing the significant impact of negation words on the overall meaning, we replace all instances of negation words (e.g., don't, can't, isn't) with "not" to ensure proper consideration of negation in a tweet.

Certain emojis serve as reliable indicators of sentiment polarity. While some words may lack inherent sentiment, the presence of emojis alongside these words can imbue the tweet with sentiment value. Consequently, we emphasize the importance of retaining emojis in the corpus, and each emoji is substituted with an alias. To address sparsity and reduce vocabulary size, we perform two straightforward operations. Firstly, we lowercase every word in our corpus. Secondly, we eliminate characters that are repeated at least three times within a word. For example, "goooood" is transformed into "good." The final preprocessing step involves stemming, a process that minimizes morphological variations by reducing words to a common root or stem. In our approach, we employed the Porter stemmer, which reduces words like "saddest," "sadness," and "sadly" to the common root "sad."

c. **Feature Engineering**

Datasets consist of a training set with 41,157 entries and a testing set with 3,798 entries, both containing two columns. Moving forward, a crucial aspect of the analysis involves understanding the unique sentiments present in the training data, namely 'Neutral,' 'Positive,' 'Extremely Negative,' 'Negative,' and 'Extremely Positive.'

```
1 train.shape,test.shape
]: ((41157, 2), (3798, 2))
```

Two pandas test and train are outlined, the first containing 3798 entries and the second with 41157 entries. Each comprises three columns: 'OriginalTweet,' holding the tweet content; 'Sentiment,' indicating the sentiment labels; and 'encoded_cat,' representing encoded categorical sentiments as int8 values. Both datasets exhibit non-null values in their respective columns, reflecting comprehensive information. The efficient use of int8 data types minimizes memory usage, making the DataFrames suitable for storage and analysis. The provided summary encapsulates key details about the dataset structures, sizes, column contents, data types, and memory efficiency.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3798 entries, 0 to 3797
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   OriginalTweet  3798 non-null   object
 1   Sentiment      3798 non-null   object
 2   encoded_cat    3798 non-null   int8
dtypes: int8(1), object(2)
memory usage: 63.2+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   OriginalTweet  41157 non-null  object
 1   Sentiment      41157 non-null  object
 2   encoded_cat    41157 non-null  int8
dtypes: int8(1), object(2)
memory usage: 683.4+ KB
```

c. Unique Sentiments in Training Data

The exploration of unique sentiments provides a foundational understanding of the emotional expressions contained within the dataset. The identified sentiments serve as key categories that play a pivotal role in subsequent analyses.

```
1 train['Sentiment'].unique()
]: array(['Neutral', 'Positive', 'Extremely Negative', 'Negative',
       'Extremely Positive'], dtype=object)
```

i. Sentiment Counts in Training and Testing Data

A detailed breakdown of sentiment counts in both the training and testing datasets reveals the distribution of sentiments across the entire dataset. This quantitative overview is essential for comprehending the prevalence of different sentiments.

```
In [5]:    1 train.Sentiment.value_counts()

Out[5]: Sentiment
        Positive             11422
        Negative              9917
        Neutral               7713
        Extremely Positive    6624
        Extremely Negative    5481
        Name: count, dtype: int64
```

```
In [6]:    1 test.Sentiment.value_counts()

Out[6]: Sentiment
        Negative             1041
        Positive              947
        Neutral               619
        Extremely Positive    599
        Extremely Negative    592
        Name: count, dtype: int64
```

ii. Encoding Sentiments

An intriguing aspect of the analysis involves encoding sentiments, as reflected in the creation of a new column named 'encoded_cat' in the training dataset. This transformation into numerical values facilitates the application of machine learning models, contributing to the effectiveness of sentiment analysis.

```
1 train['encoded_cat'] = train['Sentiment'].astype('category').cat.codes
2 train.head()
```

	OriginalTweet	Sentiment	encoded_cat
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	3
1	advice Talk to your neighbours family to excha...	Positive	4
2	Coronavirus Australia: Woolworths to give elde...	Positive	4
3	My food stock is not the only one which is emp...	Positive	4
4	Me, ready to go at supermarket during the #COV...	Extremely Negative	0

iii. Head of the Training Data with Encoded Categories

Examining the initial rows of the training dataset, which include the original tweets, associated sentiments, and newly encoded categories, provides a practical illustration of the dataset structure.

```
   1 train.encoded_cat.value_counts()

]: encoded_cat
   4    11422
   2     9917
   3     7713
   1     6624
   0     5481
   Name: count, dtype: int64
```

iv. Encoded Category Value Counts

Delving deeper, the analysis highlights the counts of each encoded sentiment category in the training dataset. This breakdown sheds light on the distribution of sentiments after encoding.

```
In [8]:    1 train.Sentiment.value_counts()

Out[8]: Sentiment
        Positive             11422
        Negative              9917
        Neutral               7713
        Extremely Positive    6624
        Extremely Negative    5481
        Name: count, dtype: int64
```

```
In [10]:    1 test.Sentiment.value_counts()

Out[10]: Sentiment
         Negative             1041
         Positive              947
         Neutral               619
         Extremely Positive    599
         Extremely Negative    592
         Name: count, dtype: int64
```

4. MODELS

Our model architecture combines a multilingual transformer with attention mechanisms finetuned for sentiment analysis. We delve into the specifics of the architecture, explaining how it accommodates multilingual data and addresses challenges unique to social media language. A detailed examination of the training process, hyperparameter tuning, and the rationale behind the chosen architecture enriches the technical understanding.

5. EVALUATIONS

In this segment, we showcase the outcomes of the models' assessment. Initially, we conducted training on the training subset, encompassing 80% of the dataset, and subsequently assessed their performance on the testing subset, which comprises 20% of the dataset.

Train and Test for original tweet

```
In [14]:  X_train=list(train['OriginalTweet'])
          X_train[:5]

Out[14]: ['@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8',
 'advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours scho
ols employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order',
 'Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/
bInCA9Vp8P',
 "My food stock is not the only one which is empty...\r\r\n\r\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYON
E if you do not take more than you need. \r\r\nStay calm, stay safe.\r\r\n\r\r\n#COVID19france #COVID_19 #COVID19 #coronavi
rus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j",
 "Me, ready to go at supermarket during the #COVID19 outbreak.\r\r\n\r\r\nNot because I'm paranoid, but because my food sto
ck is litteraly empty. The #coronavirus is a serious thing, but please, don't panic. It causes shortage...\r\r\n\r\r\n#Coro
navirusFrance #restezchezvous #StayAtHome #confinement https://t.co/usmuaLq72n"]
```

```
In [15]:  X_test=list(test['OriginalTweet'])
          X_test[:5]

Out[15]: ['TRENDING: New Yorkers encounter empty supermarket shelves (pictured, Wegmans in Brooklyn), sold-out online grocers (FoodK
ick, MaxDelivery) as #coronavirus-fearing shoppers stock up https://t.co/Gr76pcrLWh https://t.co/ivMKMsqdT1',
 "When I couldn't find hand sanitizer at Fred Meyer, I turned to #Amazon. But $114.97 for a 2 pack of Purell??!!Check out h
ow  #coronavirus concerns are driving up prices. https://t.co/ygbipBflMY",
 'Find out how you can protect yourself and loved ones from #coronavirus. ?',
 '#Panic buying hits #NewYork City as anxious shoppers stock up on food&amp;medical supplies after #healthcare worker in he
r 30s becomes #BigApple 1st confirmed #coronavirus patient OR a #Bloomberg staged event?\r\r\n\r\r\nhttps://t.co/IASiReGPC4
\r\r\n\r\r\n#QAnon #QAnon2018 #QAnon2020 \r\r\n#Election2020 #CDC https://t.co/29isZOewxu',
 '#toiletpaper #dunnypaper #coronavirus #coronavirusaustralia #CoronaVirusUpdate #Covid_19 #9News  #Corvid19 #7NewsMelb #du
nnypapergate #Costco    One week everyone buying baby milk powder the next everyone buying up toilet paper. https://t.co/Sc
ZryVvsIh']
```

6. RESULTS AND DISCUSSION

	precision	recall	f1-score	support
Extremely Negative	1.00	1.00	1.00	592
Extremely Positive	1.00	1.00	1.00	599
Negative	1.00	1.00	1.00	1041
Neutral	1.00	1.00	1.00	619
Positive	1.00	1.00	1.00	947
accuracy			1.00	3798
macro avg	1.00	1.00	1.00	3798
weighted avg	1.00	1.00	1.00	3798

The classification report outlines precision, recall, and F1-score metrics for each sentiment category. The model exhibits outstanding performance, achieving perfect scores across all metrics for each sentiment class, including 'Extremely Negative,' 'Extremely Positive,' 'Negative,' 'Neutral,' and 'Positive.' The accuracy score is also exceptionally high, reaching 100%, indicating the model's ability to accurately classify sentiments in the given dataset. The macro and weighted averages reinforce the overall excellence of the model's performance, demonstrating its robustness in handling various sentiment categories related to COVID-19. It's worth noting that achieving such perfect scores is rare in practical scenarios and suggests a highly effective sentiment analysis model in this specific context.

7. REAL-WORLD DEPLOYMENT DEMO

Objective: Sentiment Analysis for Multilingual Social Media Posts

Features: 1. Analyzing sentiments in CSV and Text files

2. Predicting accuracy using BERT models

3. Visualizing sentiment distribution

The goal of our application is to perform sentiment analysis on social media posts related to COVID-19. We support various file formats, such as CSV and Text, and provide functionalities for predicting accuracy and visualizing sentiment distribution.

Output

Components

- File Selection: Browse and load CSV or Text files
- File Type: Choose between CSV, Text, or Other
- Analyze Sentiment Button
- Predict Accuracy Button
- Result Labels
- Listbox for Sentiment Outputs
- Plot Sentiment Distribution Button
- Additional Plots for Positive, Neutral, and Negative Sentiments

Functionality

- Use the 'Browse' button to select and load CSV or Text files
- Display the loaded file path and the number of entries

- The image you sent shows the output of a sentiment analysis program for the text "Image, Faces IMG_5882.jpg, "rectanglı 1".
- The overall sentiment is 4 stars (0.3050). This means that the program has determined that the text expresses a positive sentiment, with a confidence score of 30.5%.
- The program has also analyzed the sentiment of each individual sentence in the text. The sentences "Image, Faces IMG_5882.jpg, "rectanglı 1" and "IMG_5883.jpg, "rectangles[[(1505,803 1" have been classified as positive, with sentiment scores of 0.3050 and 0.2597, respectively.
- The sentences "IMG_5884.jpg, rectangles[]" and "IMG_5885. 1" have been classified as neutral, with sentiment

scores of 0.2371 and 0.2180, respectively.

CONCLUSION

The sentiment analysis results based on COVID19 demonstrate an exceptional performance of the model. With perfect precision, recall, and F1score across all sentiment categories, including 'Extremely Negative,' 'Extremely Positive,' 'Negative,' 'Neutral,' and 'Positive,' the model exhibits an unparalleled accuracy of 100%. The macro and weighted averages further affirm the model's robustness, indicating its ability to effectively handle diverse sentiment expressions related to the pandemic. While achieving such flawless scores is uncommon, it underscores the effectiveness of the sentiment analysis model in accurately capturing and categorizing sentiments within the provided dataset. This outcome showcases the model's reliability and suitability for sentiment analysis tasks in the context of COVID-19.The Sentiment Analysis GUI combines the power of BERT models with an intuitive interface. Users can perform accurate sentiment analysis, predict accuracy, and gain insights through visually appealing plots, making it a comprehensive tool for social media sentiment analysis.

REFERENCES

[1]. Banea, C., Mihalcea, R., & Wiebe, J. (2010). Multilingual subjectivity: Are more languages better? Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).

[2]. Bautin, M., Vijayarenu, L., & Skiena, S. (2008). International sentiment analysis for news and blogs. ICWSM.

[3]. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

[4]. Kim, S.-M., & Hovy, E. (2006). Identifying and analyzing judgment opinions. Proceedings of the human language technology conference of the NAACL, main conference.

[5]. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer.

[6]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[7]. Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130137.

[8]. Portuguese Tweets for Sentiment Analysis. (2019). Portuguese Tweets.

[9]. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis:

tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

[10]. Saad, M. (2020). Arabic Sentiment Twitter Corpus. Arabic Tweets.

[11]. Shin, B., Lee, T., & Choi, J. D. (2016). Lexicon integrated CNN models with attention for sentiment analysis. arXiv preprint arXiv:1610.06272.

[12]. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[13]. Statista. (2019). Internet: most common languages online 2020 | Statista. [14]. Taboada, M., Brooke, J.,
Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[15]. Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2016). Dimensional sentiment analysis using a regional CNN-LSTM model. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers).

CONTRIBUTORS

1. Akhila Awoshetty
2. Mohammed Khaja mujahiddin
3. Chandramohan Chandanakumar