TWITTER SENTIMENT ANALYSIS

Trending subject

Supervised by
dr. Diaa
With the help of
Engineer . Mohamed

- Mahmoud amer 41810138
- Menna amr 41810100
- Shrouk Yasser 41810072
- Nada elsayed 41810306

Twitter Trend Analyzer

Sentiment in the modern age , positive negative , netural analysis

***Abstract*—This research takes the social network Twitter as its content matter due to its significant uniqueness of not being a visually based social network that is prevalent today. Twitter has been known for its straight message transmissions from the sender to the receiver. On Twitter everyone can interact with everyone from celebrities and politicians to farmers and students.**

Keywords—twitter, analyzer, NLP ,SDG

I. WHY TWITTER

II. This research takes the social network Twitter as its content matter due to its significant uniqueness of not being a visually based social network that is prevalent today. Twitter has been known for its straight message transmissions from the sender to the receiver. On Twitter everyone can interact with everyone from celebrities and politicians to farmers and students. This research aims to answer the concept of correlations and trends between Twitter users and their social network scope. Being branded as the "social network for smarter people", Twitter is relevant to the intellectual property that composes the electronic knowledge base. Twitter is a witty and smart social network; users have 140 characters or less to compose their message EASE OF USEENT ANALYSIS

I. Why twitter

Twitter allows businesses to engage personally with consumers. That's why sentiment analysis has become a key instrument in social media marketing strategies. Sentiment analysis is a tool that automatically monitors emotions in conversations on social media platforms..

I. Twitter is better optimized for this task

Twitter is a witty and smart social network; users have 140 characters or less to compose their message . Twitter offers distinctive ways of communication that are mainly text based. A Twitter user interested in the statuses of another user signs up to be a "follower . Twitter has been a source for event evacuation like the Haiti earthquakes and people have used the medium to save many human

Stages :

1. **1. Gather Twitter Data**

- It's important that your Twitter data is representative of what you're trying to find out because you'll use it to:
- Train your sentiment analysis model
- Test how your model performs on Twitter data

2. **It's important that your Twitter data is representative of what you're trying to find out because you'll use it to:**

- Train your sentiment analysis model
- Test how your model performs on Twitter data

3. **2. Prepare Your Data**

- Once you've gathered the tweets you need for your sentiment analysis, you'll need to prepare your data. Social media data is unstructured and needs to be cleaned before using it to train a sentiment analysis model – good quality data will lead to more accurate results
- Preprocessing a Twitter dataset involves a series of tasks like removing all types of irrelevant information like emojis, special characters, and extra blank spaces. It can also involve making format improvements, delete duplicate tweets, or tweets that are shorter than three characters.

4. **3. Create a Twitter Sentiment Analysis Model**

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning.

SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features.

- The advantages of Stochastic Gradient Descent are:

➢ Efficiency.
➢ Ease of implementation (lots of opportunities for code tuning).

- The disadvantages of Stochastic Gradient Descent include:

  - SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations.
  - SGD is sensitive to feature scaling

- Keywords :

1. **Data analysis**: Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.
2. **Tokenization:** Tokenization is the process of turning a meaningful piece of data, such as an account number, into a random string of characters called a token that has no meaningful value if breached. Tokens serve as reference to the original data, but cannot be used to guess those values. *Units*

3. **computer vader scores** : Compound VADER scores for analyzing sentiment The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive)
4. **pd.get_dummies**: Convert categorical variable into dummy/indicator variables.
5. **Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.
6. **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good, Precision = TP/TP+FP
7. **Recall** (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it's above 0.5, Recall = TP/TP+FN
8. **F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

I. TWITTER SENTIMENTAL ANALYSIS

Social networks is a rich platform to learn about people's opinion and sentiment regarding different topics as they can communicate and share their opinion actively on social medias including Facebook and Twitter. There are different opinion oriented information gathering systems which aim to extract people's opinion regarding different topics.

The sentiment-aware systems these days have many applications from business to social sciences. Since social networks, especially Twitter, contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text [15]. Some of the top abbreviations are FB for Facebook, B4 for before, OMG for oh my god and so on. Therefore sentimental analysis for short texts like Twitter's posts is challenging [8].

IV. DESIGN AND IMPLEMENTATION

This technical paper reports the implementation of the Twitter sentiment analysis, by utilizing the APIs provided by Twitter itself. There are great works and tools focusing on text mining on social networks. In this project the wealth of available libraries has been used.

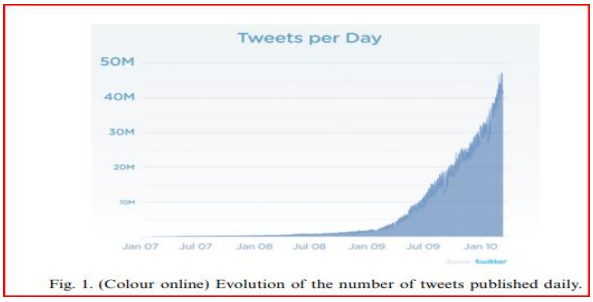


Fig. 1. (Colour online) Evolution of the number of tweets published daily.

Algorithm 1 Extract Twitter sentiment

```
procedure TWITTER-CONNECTION()
    consumer − key =' xxxxxxxxx'
    consumer − secret =' xxxxxxxx'
    access − token =' xxxxxxxx'
    access − token − secret =' xxxxxxxxxx'
    self.auth = OAuthHandler(consumer − key, consumer − secret)
    self.auth.set − access − token(access − token, access − token − secret)
    self.api = tweepy.API(self.auth)
end procedure

procedure TWEET-CLEANING(t)
    tweet = t.remove − Stop − words
    Return tweet
end procedure

procedure TWEET-CLASSIFICATION(t)
    t = Tweet − Cleaning(t)
    tweet − polarity = t.sentiment.polarity
    tweet − polarity
end procedure

procedure GET-TWEETS(q, count)
    fetched − tweets = self.api.search(q = query, count = count)
    Return fetched − tweets
end procedure

procedure MAIN()
    st = SentimentalTwitter()
    tweets = st.fetch − tweets(query =' politics', count = 300)
    PositiveTweets = tweetsthatsentiment =' positive'
    NegativeTweets = tweetsthatsentiment =' negative'

    for tweet t in PositiveTweets do
        print(t)
    end for
    for tweet t in NegativeTweets do
        print(t)
    end for
end procedure
```

Figure 2. Extract Sentiment from Twitter using API

III. IMPLEMENTATION

In this paper, we used python to implement sentiment analysis. Some packages have utilized including nltk , tensorflow , we can install the required libraries by following commands :

- Pip install nltk
- Pip install tensorflow –v

The second step is importing the dataframe and assigning it to a new df , so change we make don't affect the original dataset :

df0 = pd.read_csv('2021_May_twitter_trending_data.csv')
df0

The textblob is a python library for text processing and it uses NLTK for natural language processing [6].

A. *Convert tweet text into a sequence of words*

```
def tweet_to_words(tweet):
    ''' Convert tweet text into a sequence of words '''

    # convert to lowercase
    text = tweet.lower()
    # remove non letters
    text = re.sub(r"[^a-zA-Z0-9]", " ", text)
    # tokenize
    words = text.split()
    # remove stopwords
    words = [w for w in words if w not in stopwords.words("english")]
    # apply stemming
    words = [PorterStemmer().stem(w) for w in words]
    # return list
    return words
```

B. *train the model*

```
history = model.fit(X_train,y_train,validation_data=(X_val, y_val),batch_size=64,epochs=epochs,verbose=1)
```

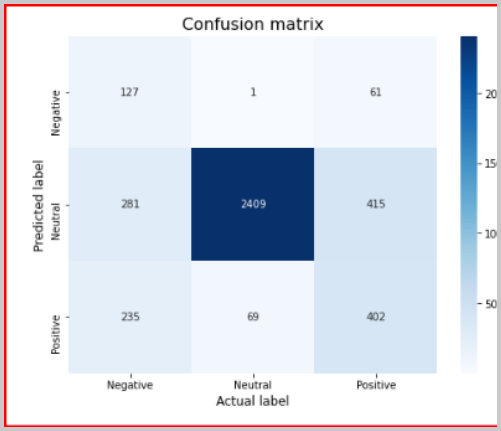
C. *Evaluate model on the test set*

```
Accuracy  : 0.7345
Precision : 0.8132
Recall    : 0.6062
F1 Score  : 0.6946
```

The Evaluation Matrix

D. Confusion matrix

Compute confusion matrix to evaluate the accuracy of a classification, A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. ... The rows



confusion matrix 1

represent the predicted values of the target

Query	Positive	Negative	Neutral
Tweet_text	12284	4362	3354

CONCLUSION

In this technical paper, we discussed the importance of social network analysis and its applications in different areas. We focused on Twitter as and have implemented the python program to implement sentimental analysis. We showed the results on different daily topics. We realized that the neutral sentiments are significantly high which shows there is a need to improve Twitter sentiment analysis.

REFERENCES

[1] Boguslavsky, I. (2017). Semantic Descriptions for a Text Understanding System. In Computational Linguistics and Intellectual Technologies.Papers from the Annual International Conference "Dialogue"(2017)(pp.14-28). J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010). K. Elissa, "Title of paper if known," unpublished.

[3] Scott, J. (2011). Social network analysis: developments, advances, and prospects. *Social network analysis and mining*, *1*(1), 21-26.

[4] Statista, 2017, https://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/

[5] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115-120). Association for Computational Linguistics.

[6] TextBlob, 2017, https://textblob.readthedocs.io/en/dev/

[7] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 135.

[8] Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

[9] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*(pp. 347-354). Association for Computational Linguistics.

[10] https://www.researchgate.net/publication/234052505_Sentiment_analysis_in_Twitter/link/02e7e526116fcc48fe000000/download

[11] https://scikit-learn.org/stable/modules/sgd.html

[12]https://towardsdatascience.com/a-beginners-guide-to-match-any-pattern-using-regular-expressions-in-r-fd477ce4714c

[13]https://www.ieee.org/searchresults/index.html?q=machine+learning+#gsc.tab=0&gsc.q=machine%20learning%20&gsc.page=1

[14] https://arxiv.org/abs/1811.00659

[15] https://osf.io/preprints/socarxiv/8numc/

[16] PYTHON REGULAR EXPRESSIONS The ultimate crash course to learn Python Regular Expressions FAST! Pdf

[17] minning the social web by Matthew A. Russell

[18] Applied Text Analysis with Python by Benjamin Bengfort , Rebecca Bilbord & Tony ojeda

[19] Natural Language processing with PyTorch by Delip Rao& Brian McMahan

[20] Natural Language Processing Recipes Unlocking Text Data with Machine Learning and Deep Learning using python