Twitter Sentiment Analysis with Neural Networks: K-pop Industry

By Francis Jin

Topic

This project seeks to gain insight into trends into the ever-evolving and increasingly popular Korean pop music industry, deemed "K-pop", which has risen to one of the top musical genres in the world. By using facets of Natural Language processing and machine learning to conduct Sentiment Analysis on Twitter data, the analysis measures trends in popularity as well as variations amongst different K-pop groups.

Ultimately, the project attempts to test the hypothesis that *Twitter sentiment is a useful signal for the overall popularity and success of a musical artist group.*

Metric

The metric we will use to test this hypothesis is the overall score the model achieves, when comparing its rankings of the top 10 K-pop groups right now to the ranking order done by Koreaboo, one of the largest online K-pop content media platforms (**https://www.koreaboo.com/news/top-30-popular-kpop-groups-korea-right-now/**).

A strong overlap will signify the model's ability to predict a K-pop groups popularity using Twitter sentiment, and overall popularity is known to be strongly correlated with the overall success of a musical group.

Methods

Search set

Using Tweepy API, a query function retrieves the last 3,000 tweets based on the following artist names as keywords: 'BTS', 'TWICE', 'Red Velvet', 'ITZY', 'BLACKPINK', 'Mamamoo', 'Oh My Girl', 'Girls' Generation', 'IZ*ONE', 'Lovelyz'.

Note that some groups with names such as "red velvet" have other common use linguistic use cases such as "red velvet cake", and are thus subject to more noise in the data. We solve this by re-running the Search Set on those groups with "red velvet kpop", to target only the tweets pointing to the musical group.

Training Set

Built with labeled data using a corpus file with ID keys to 5000 sentiment-labeled tweets, which we grab through the Twitter API, without saving any additional information as to comply with the Twitter Developer API usage rules. Then it is written to CSV and mapped to numeric values 0 for "negative", 1 for "neutral", 2 for "positive", and 4 for "irrelevant".

```
Value counts for each sentiment label
1    2333
4    1689
0     572
2     519
Name: Sentiment, dtype: int64
```

Pre-Processing

Commonly in Natural language processing endeavors, text must be processed to be suitable for modeling. Here we use the following Python libraries such as nltk, re, spacy, to edit our texts to convert to lowercase, remove whitespace, remove personal pronouns, remove URLs and the # sign in hashtags, and simplify repeated characters. We then filter out the entries labeled “4” (irrelevant).

Feature Selection

We fit and transform the data with a TFIDF transformer, term frequency–inverse document frequency, which weights words based on importance in a document.

Then we use a Glove embedding, an unsupervised learning algorithm for distributed word representation, to create vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity.

Models

Keras Neural Network

A sequential neural network from Keras library is used for multi-label classification on the sentiment labels 0, 1, and 2. A single inner layer, early stopping, and Dropout are used in the neural network with a SoftMax activation layer in the output, and loss set to categorical cross-entropy.

We achieve scores of about 80% and 80% accuracy for Training and Validation respectively.

```
Training accuracy: 82.02839493751526
Validation accuracy: 77.73722410202026 for the Glove embeddings.
```

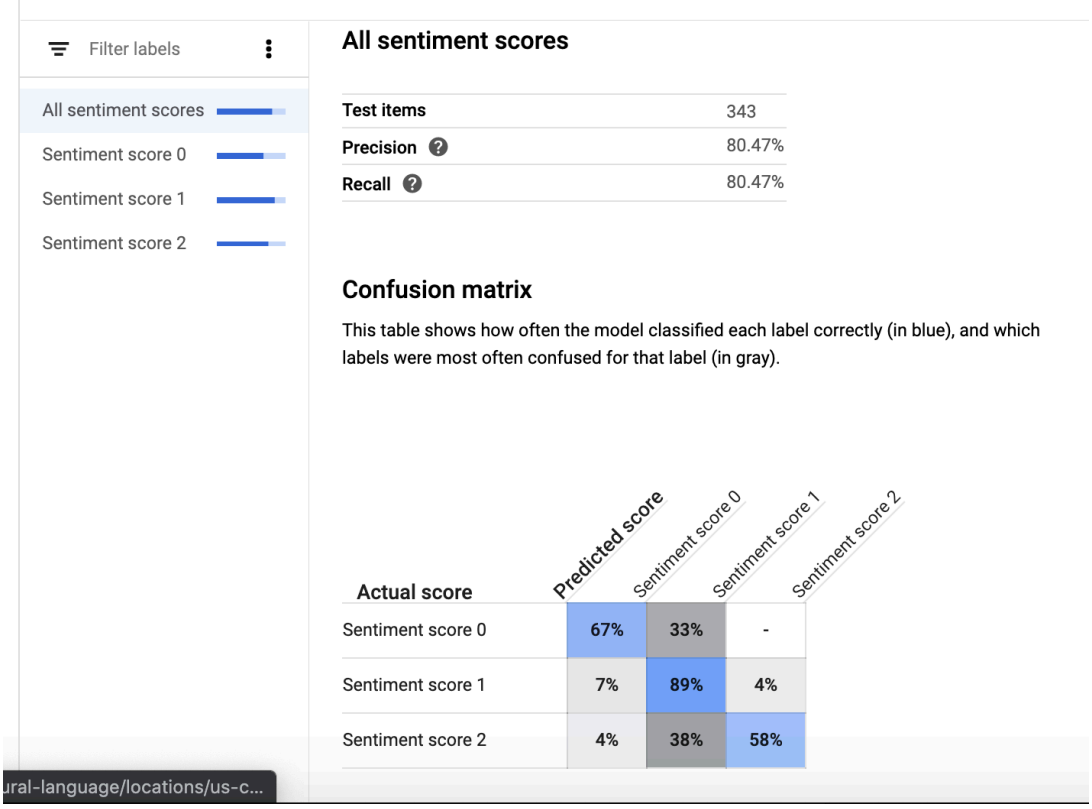
After training the model and getting decent accuracy results, we use it to label the Search Set and generate the proportion of positive, negative, and neutral classified tweets on our keyword in question. In this example, “blackpink”. Strongly positive!

```
Percentage of Negative(0), Neutral(1), and Positive(2) Tweets for keyword: blackpink from a total of 10000 tweets
2    93.41
1     3.47
0     3.12
Name: glovelabel, dtype: float64
```

AutoML

Google Cloud Platform’s AutoML service is useful for evaluating the viability of a model for NLP, and depending on the use case can generate very sufficient ready-to-deploy machine learning models. In our case, the training set is published to AutoML to train an NLP model for Sentiment Analysis for the labels: 0 for negative, 1 for neutral, 2 for positive.

We see the results of AutoML closely mirror that of my original Sequential neural network! Precision and Recall scores of 80.47%, Confusion Matrix below.



Results

We rank each K-pop group in order based on ratio of positive to negative tweets to create Sentiment Index rankings, and compare to Koreaboo Magazine Rankings. Most groups landed very close to real ranking! This supports our original hypothesis that Twitter sentiment is indicative of overall popularity.

Artist	Sentiment Index	Sentiment Rankings	Koreaboo Rankings	Ranking Diff
0 TWICE	266.630000	1	2	+1
1 Lovelyz	183.930000	2	10	+8
2 BTS	112.110000	3	1	-2
3 Red Velvet	78.990000	4	3	-1
4 Blackpink	62.900000	5	5	+0
5 IZ*ONE	31.350000	6	9	+3
6 Girls' Generation	25.230000	7	8	+1
7 Mamamoo	21.350000	8	6	-2
8 ITZY	17.710000	9	4	-5
9 Oh My Girl	7.160000	10	7	-3

*Note that every single group received an **overall positive sentiment rating**, not surprising given that our subjects are Pop music groups in the entertainment industry! (As opposed to for example, controversial political topics).

This insight can be used by players in the media and entertainment industry to measure sentiment, social climate, track popularity over time, and inform numerous business practices such as marketing, promotion, and strategy.