KDD & Feature Engineering

Group Name - Abraca-Data

Table of Contents

Introduction

With the advent of the Internet and web technologies, people have changed their way of consuming news in their life. Traditional physical newspapers have been replaced by online versions like online news and videos. Readers are more inclined to use online sources of news mainly due to two key features: interactivity and immediacy. Today, people want to consume as much news, from as many sources, as they possibly can, on matters that are important to them or matters that catch their attention.

Interactivity refers to the inherent tendency depicted by the masses that makes them consume news of their interest where as Immediacy is a feature that represents the need of people to be informed about news with no delay in time. Online news expresses opinions regarding news entities, which may comprise of people, places or even things, while reporting on events that have recently occurred.

Sentiment Analysis and Opinion Mining is a way of finding out the polarity or strength of the opinion (positive, negative or neutral) that is expressed in written text, in the case of news articles. The work described in this project can be a module in a opinion mining/ opinion monitoring system. This module would extract sentiment from natural language and provide it as an input to that larger system.

There are two popular approaches that are utilized to automate the process of sentiment analysis.The first process makes use of a lexicon of weighted words and the second process is based on approaches of machine learning. Lexicon based methods use a word stock dictionary with opinion words and match given set of words in a text for finding polarity. As opposed to machine learning methods, this approach does not need to preprocess data not does it have to train a classifier. The results of both of these approaches are reported to the voting algorithm. Due

to the time and resources limitation of this project, the problem is open-ended. It gives the foundation of the basic ideas of how to solve the problem using natural language processing (NLP) techniques, but does not describe a complete opinion mining and monitoring system. This work gives an example of how to parse information and link that information with the current news topics.

Dataset

Our articles are taken from The New York Times , CNBC and Twitter Timeline. Our articles corpus contains every article published from Jan 2019 to Nov 2020. Each article is annotated with published date, article link and meta article information describing the content of the article. We are using twitter developer api program and news api to collect and store data in database.

Dataset Collection, Preprocessing , Transformation and Analysis

The methodology comprised of 3 steps, starting with data collection.

- **Data Collection -** News Articles are collected on a daily basis using twitter developer api and newyork times api. Collected data are first parsed through python package module and then stored in a nosql database in key-value pairs. We parse and extract following fields from the news article **- source name, article title, article authors, article published date, article text, images link, videos link, article summary, article keywords** and **article url**. We also collects and updates the number of followers the article authors have on twitter .
- **Data Preprocessing and Transformation-** Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete. Preprocessing is a necessary step to clean text (lessen noise of text) and to reduce inconsistencies from it so that this cleansed data can more effectively be utilized in text mining or sentiment analysis task . Data preprocessing resolves such issues and makes datasets more complete and efficient to perform data analysis. we use NLTK, Spacy, and some regular expressions to preprocess the news articles, process involved are **Removing (punctuations, Quotes, Brackets, Currency Chars and Digits) , Removing URLs, Removing Stop words, Lower casing , Sentence Segmentation, Tokenization, Stemming and Lemmatization, Named Entity recognition**
- **Analysis -** Sentiment analysis can generally be carried out using **supervised or unsupervised approaches**. A supervised approach comprises of a set of labeled training data that is used to build a classification model with the intent of using this model to classify new data for which labels are not present.
  Unsupervised or Lexicon-based approaches to sentiment analysis do not require any training data. In this approach, the sentiments conveyed by a word are inferred on grounds of the polarity of the word. In case of a sentence or a document, the polarities of the individual words that compose the document collectively convey the sentiment of the sentence or the document. Thus the polarity of a sentence is the accumulative total (sum) of polarities of the individual words (or phrases) in the sentence. This approach utilizes some predefined lists of words such that each word in the list is associated with a specific sentiment. Further this approach can use the following methods:

  1. **Dictionary-based methods**: in these methods lexicon dictionary is used in order to find out the positive opinion words and negative opinion words.
  2. **Corpus-based methods**: in these methods large corpus of words is used and based on syntactic patterns other opinion words can be found within the context.

     Sentiment analysis can be done on document level, sentence level, word level or phrase level. we will be using **Text Blob and VADER** for Lexicon-based approaches to sentiment analysis and **Embeddings & Transformers** for Machine learning approach.

     We use scores to track two trends over time:

     • Polarity: Is the sentiment associated with the entity positive or negative?

• Subjectivity: How much sentiment (of any polarity) does the entity garner?

Subjectivity indicates proportion of sentiment to frequency of occurrence, while polarity indicates percentage of positive sentiment references among total sentiment references.

We focus first on polarity. We evaluate world polarity using sentiment data for all entities for the entire time period: **world polarity = positive sentiment references / total sentiment references**

entity polarity = positive sentiment references/ total sentiment references

Subjectivity scores - The subjectivity time series reflects the amount of sentiment an entity is associated with, regardless of whether the sentiment is positive or negative. Reading all news text over a period of time and counting sentiment in it gives a measure of the average subjectivity levels of the world.

We evaluate world subjectivity using sentiment data for all entities for the entire time period:

world subjectivity = total sentiment references /total references

entity subjectivity = total sentiment references /total references

Evaluation Metrics

To evaluate the performance of algorithms for sentimental analysis and other nlp related tasks, various evaluation metrics have been used. Most existing approaches as a classification problem that predicts whether a news article is positive or negative:

• True Positive (TP)

• True Negative (TN)

• False Negative (FN)

• False Positive (FP)

By formulating this as a classification problem, we can define following metrics,

1. Precision = |T P| / |T P| + |F P|
2. Recall = |T P| / |T P| + |F N|
3. F1 = 2 * (Precision · Recall)/ Precision + Recall
4. Accuracy = |T P| + |T N| / |T P| + |T N| + |F P| + |F N|

   These metrics are commonly used in the machine learning community and enable us to evaluate the performance of a classifier from different perspectives.

Feature Extraction

News Content Features describe the meta information related to a piece of news.

A list of representative news content attributes are listed below:

- Source Name: Publisher of the news article
- Article Title: Short title text that aims to catch the attention of readers and describes the main topic of the article
- Article Authors : Name of the authors who published the news.
- Article Published Date : Date on which article is published.
- Body Text: Main text that elaborates the details of the news story; there is usually a major claim that is

specifically highlighted and that shapes the angle of the publisher

- Images Link: Part of the image body content of a news article that provides visual cues to frame the story
- Video Link: Part of the video body content of a news article that provides visual cues to frame the story
- Article Summary : Summary created from the article body , keeping the relavant information.
- Article keywords: Keywords related to the content of the article body.
- Article Url: Published Article URL