Introduction to Data Science and Analytics

Project Title
Social Media Sentiment Analysis

1. Project

Sentiment analysis, a sub-field of Natural Language Processing, is one of the most popular topics and research fields in data science. We will be working on social media sentiment analysis. We aim to be able to classify tweets, reviews and comments from social media as positive, negative or neutral.

The most important point of our project is data mining to collect a large amount of data from several sources. For this purpose, we found open source datasets such as Sentiment140 [1] and many others. Besides, we would like to collect our own data from social media and expand our dataset. We also found some tools and APIs to retrieve new data. TWINT - Twitter Intelligence Tool [2] is an advanced Twitter scraping tool written in Python that allows for scraping tweets. One of the most important features of TWINT is that there is no need to use Twitter Developer API. This feature allows collecting data with no rate limitations.

Most of the open-source datasets that we found on the internet are properly labeled and structured. Data collected by ourselves need to be properly labeled. Then, we will go through the cleaning, preprocessing and separation of test and training data steps.

We searched for some tools for our project and found some popular and powerful open-source NLP frameworks in Python. We will probably use Natural Language Toolkit (NLTK) [3]. It comes with all the pieces you need to get started on sentiment analysis.

2. Data

Dataset : Sentiment140

This dataset contains 1,600,000 tweets extracted using the twitter api. The tweets have been classified from 0 (negative) to 4 (positive). The dataset contains 6 fields which are target as integer, ids as integer, date as date, flag as string, user as string and text as string.These 6 fields are shown below.

- target: The polarity of the tweet (0 - negative, 2 - neutral, 4 - positive)
- ids: The id of the tweet.
- date: The date of the tweet.
- flag: The query. If there is no query, then this value is NO_QUERY.
- user: The user that tweeted.
- text: The text of the tweet

Target	Ids	Date	Flag	User	Text
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOn	@switchfoot http://twitpic.com/2y1zl - Awww, t
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by text
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Mana
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollyw	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?

Figure 1. A sample from the dataset

Dataset : The dataset consists of tweets or comments and sentimental labels. We will be using Twitter Developer API or a web scraping library called Twint - sample output shown in Figure 2 - to create our own dataset from twitter. This sample has 3 fields which are id as integer, username as string and tweet as string.

	id	username	tweet
0	1377434193827860480	joebiden	The American Jobs Plan is the largest American...
1	1377414548097880068	joebiden	Under the American Jobs Plan, 100% of our nati...
2	1377385089340829700	joebiden	Wall Street didn't build this country—the grea...
3	1377367815120957445	joebiden	Delivering for the American people is what the...
4	1377363181387980800	joebiden	The American Jobs Plan is a once-in-a-generati...
5	1377361641491300354	joebiden	Millions of Americans lost their jobs last yea...
6	1371117123859316736	joebiden	It matters whether you continue to wear a mask...
7	1370792386280972289	joebiden	The American Rescue Plan means a $7,000 check ...
8	1370411955173912576	joebiden	85% of American households will get direct che...
9	1366417985649442821	joebiden	A campaign for everyone who's been knocked dow...

Figure 2. Joe Biden's tweets collected using TWINT.

If we use Twitter Developer API or Twint to create our data set, we need to label the tweets to prepare such a supervised dataset.

Dataset : Twitter.csv

This dataset is a supervised dataset which includes tweets. Twitter.csv Dataset has around 163,000 tweets along with sentiment labels samples shown in Figure 3. This dataset has 3 fields which are id as integer, tweet as string and label as integer. Reddit.csv dataset has around 37,000 comments along with its sentimental label.

0	when modi promised “minimum government maximum...	-1.0
1	talk all the nonsense and continue all the dra...	0.0
2	what did just say vote for modi welcome bjp t...	1.0
3	asking his supporters prefix chowkidar their n...	1.0
4	answer who among these the most powerful world...	1.0
5	kiya tho refresh maarkefir comment karo	0.0
6	surat women perform yagna seeks divine grace f...	0.0
7	this comes from cabinet which has scholars lik...	0.0
8	with upcoming election india saga going import...	1.0
9	gandhi was gay does modi	1.0
10	things like demonetisation gst goods and servi...	1.0
11	hope tuthukudi people would prefer honest well...	1.0
12	calm waters wheres the modi wave	1.0
13	one vote can make all the difference anil kapo...	0.0
14	one vote can make all the difference anil kapo...	0.0
15	vote such party and leadershipwho can take fas...	-1.0
16	vote modi who has not created jobs	0.0
17	through our vote ensure govt need and deserve ...	0.0
18	dont play with the words was talking about the...	1.0
19	didn’ write chowkidar does mean ’ anti modi tr...	-1.0
20	was the one who recently said that people who ...	1.0

Figure 3. Labels indicate that 1 positive, 0 neutral and -1 negative comment.

3. References

[1] Sentiment140, http://help.sentiment140.com/home

[2] TWINT, https://github.com/twintproject/twint

[3] Natural Language Toolkit, https://www.nltk.org/