DATA MINING AND SOCIAL WEB SEMANTICS

Ms. S. Abinayaa
Assistant Professor (OG)
SRM IST
Tamilnadu

Sonali panchbhai
B.Tech Student
SRM IST
Tamilnadu

Atul Anand
B.Tech Student
SRM IST
Tamilnadu

Khushi Sureka
B.Tech Student
SRM IST
Tamilnadu

ABSTRACT

Hashtags, originally introduced in Twitter, are now becoming the most used way to tag short messages in social networks since this facilitates subsequent search, classification and clustering over those messages. However, extracting information from hashtags is difficult because their composition is not constrained by any (linguistic) rule and they usually appear in short and poorly written messages which are difficult to analyse with classic IR techniques. In this project we address two challenging problems regarding the "meaning of hashtags"— namely, hashtag relatedness and hashtag classification — and we provide two main contributions. First we build a novel graph upon hashtags and (Wikipedia) entities drawn from the tweets by means of topic annotators (such as TagME); this graph will allow us to model in an efficacious way not only classic co-occurrences but also semantic relatedness among hashtags and entities, or between entities themselves. Based on this graph, we design algorithms that significantly improve state-of-the-art results upon known publicly available datasets. The second contribution is the construction and the public release to the research community of two new datasets: the former is a new dataset for hashtag relatedness, the latter is a dataset for hashtag classification that is up to two orders of magnitude larger than the existing ones. These datasets will be used to show the robustness and efficacy of our approaches, showing improvements in F1 up to two-digits in percentage (absolute).

Keywords :- *twitter, sentiment analysis, graph.*

INTRODUCTION

As internet is growing bigger, its horizons are becoming wider. Social Media and Micro blogging platforms like Facebook, Twitter, Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgements, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analysing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. The intention is to gain an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative or neutral labels. Many people use social media sites for networking with other people and to stay up-to-date with news and current events. The sites like Twitter, Facebook offer a platform to people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss about the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, rate about the performance of not only any movie but about other products and to know about whether it will be a success or not. Therefore, sentiment analysis has wide applications and include emotion mining, polarity, classification and influence analysis. Twitter is an online networking site driven by tweets which are 140 character limited messages. Thus, the character limit enforces the use of hashtags for text classification. Currently around 6500 tweets are published per second, which results in approximately 561.6

million tweets per day. These streams of tweets are generally noisy reflecting multi topic, changing attitudes information in unfiltered and unstructured format. Twitter sentiment analysis involves the use of natural language processing to extract, identify to characterize the sentiment content. Sentiment Analysis is often carried out at two levels:

1) coarse level
2) fine level

In coarse level, the analysis of entire documents is done while in fine level, the analysis of attributes is done. The sentiments present in the text
are of two types: Direct and Comparative. In comparative sentiments, the comparison of objects in the same sentence is involved while in direct sentiments, objects are independent of one another in the same sentence.

2. SYSTEM ANALYSIS

Our project involves the analysing of real time tweets. The objective of our case study is to find the polarity of the words (in tweets) retrieved. Each step in the framework involves several sub-tasks.

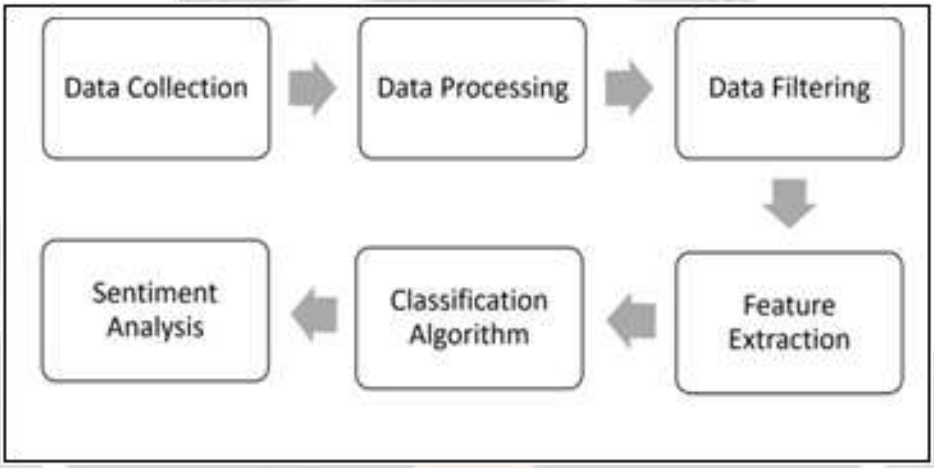


Fig. 2 Framework for Twitter Analysis

2.1 DATA COLLECTION

Data in the form of raw tweets is retrieved by using the Scala library "Twitter4j" which provides a package for real time twitter streaming API. The API requires us to register a developer account with Twitter and fill in parameters such as consumerKey, consumerSecret, accessTokenaccess and TokenSecret. This API allows to get all random tweets or filter data by using keywords. Filters supports to retrieve tweets which match a specific criterion defined by the developer. We used this to retrieve tweets related to specific keywords which are taken as input from users. Initially, we set at least set an application name and mode. We execute the program in local mode instead of cluster. Then, input array of keywords is provided as an argument to Streaming Context "ssc"using "sc" where "sc" is spark context.

For example, on inputting multiple keywords like, 'Canada', 'Trump', 'Toronto', the output we obtained from 15 seconds' window time was the live stream of tweets associated with these keywords. Only caveat of using filters is that famous keywords like "India" have more tweets compared to niche words like "Focusrite" which makes it difficult to get data for niche specific keywords.

2.2 DATA PROCESSING

Data processing involves Tokenization which is the process of splitting the tweets into individual words called tokens. Tokens can be split using whitespace or punctuation characters. It can be unigram or bigram depending on the classification model used. The bag-of-words model is one of the most extensively used model for classification. It is based on the fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest way to incorporate this model in our project is by using unigrams as features. It is just a collection of individual words in the text to be classified, so, we split each tweet using

whitespace. For example, the tweet “Met aziz today !!” is split from each whitespace as follows. {“Met Aziz !!” } The next step in data processing is normalization by conversion of tweet into lowercase. Tweets are normalized by converting it to lowercase which makes its comparison with a dictionary easier.

2.3 DATA FILTERING

A tweet acquired after data processing still has a portion of raw information in it which we may or may not find useful for our application. Thus, these tweets are further filtered by removing stop words, numbers and punctuations.

- Stop words: For example, tweets contain stop words which are extremely common words like “is”, “am”, “are” and holds no additional information. These words serve no purpose and this feature is implemented using a list stored in stopfile.dat. We then compare each word in a tweet with this list and delete the words matching the stop list.

- Removing non-alphabetical characters: Symbols such as “#@” and numbers hold no relevance in case of sentiment analysis and are removed using pattern matching. Regular expressions are used to match alphabetical characters only and rest are ignored. This helps to reduce the clutter from the twitter stream.

- Stemming: It is the process of reducing derived words to their roots. Example includes words like “fish” which has same roots as “fishing” and “fishes”. The library to use stemming is Stanford NLP which also provides various algorithms such as porter stemming. In our case, we have not employed any stemming algorithm due to time constraints.

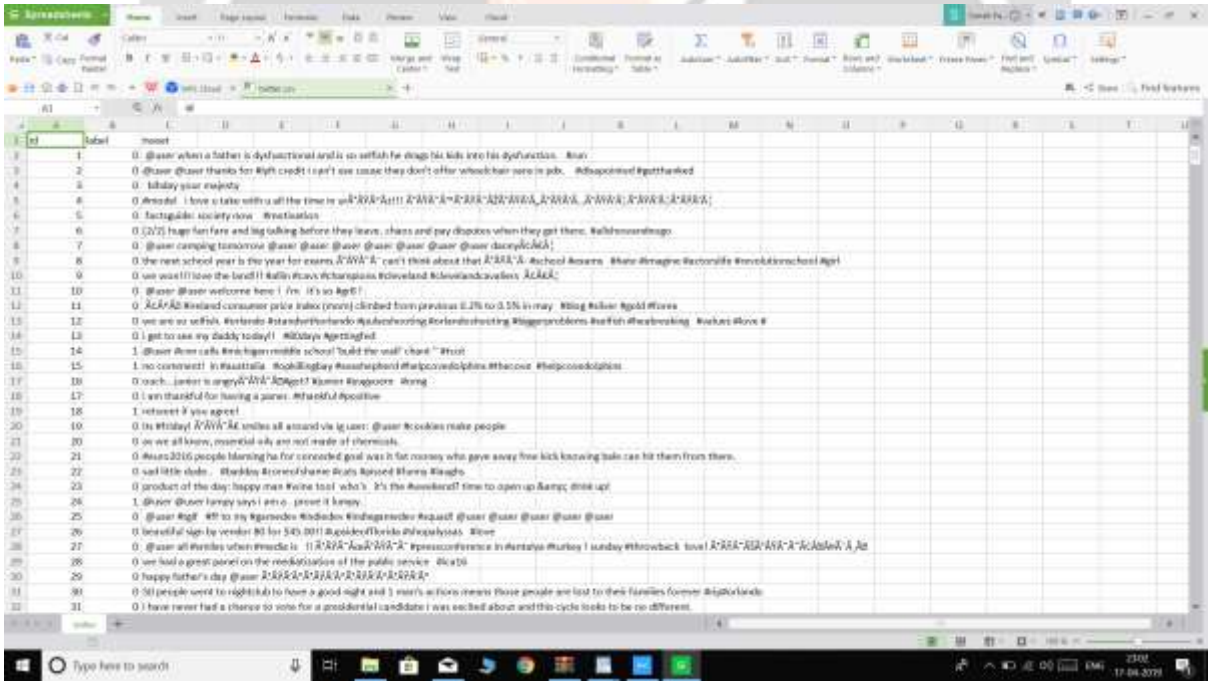
APPENDIX I (SAMPLE CODE)

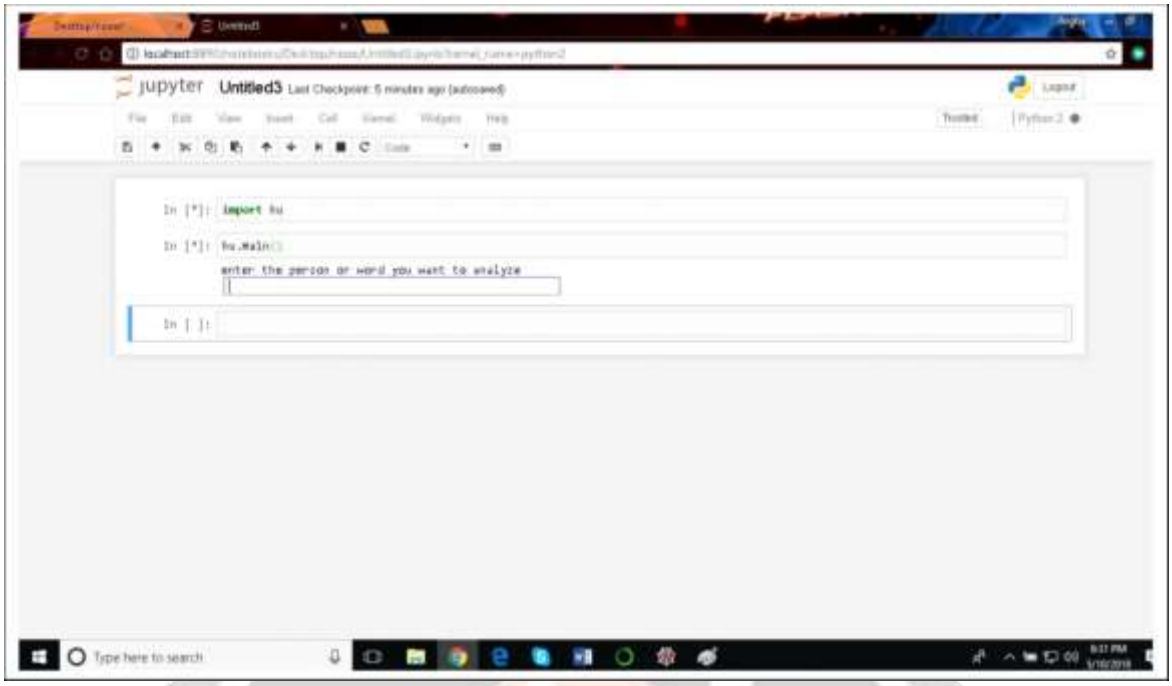
```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 1,
   "metadata": {},
   "outputs": [],
   "source": [
    "import numpy as np\n",
    "import pandas as pd\n",
    "import matplotlib.pyplot as plt"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 3,
   "metadata": {},
   "outputs": [],
   "source": [
    "train = pd.read_csv('twitter.csv')\n",
    "test = pd.read_csv('youtube.csv')\n",
    "train_test = pd.read_csv('flickr.csv')"
   ]
```

```
},
{
 "cell_type": "code",
 "execution_count": 9,
 "metadata": {},
 "outputs": [
  {
   "data": {
    "text/plain": [
     "<bound method NDFrame.head of           id  label                                              tweet\n",
     "0         1      0   @user when a father is dysfunctional and is s...\n",
     "1         2      0  @user @user thanks for #lyft credit i can't us...\n",
     "2         3      0                                bihday your majesty\n",
     "3         4      0  #model   i love u take with u all the time in ...\n",
     "4         5      0             factsguide: society now    #motivation\n",
     "5         6      0  [2/2] huge fan fare and big talking before the...\n",
     "6         7      0   @user camping tomorrow @user @user @user @use...\n",
     "7         8      0  the next school year is the year for exams.Ã°ÂŸÂ˜...\n",
     "8         9      0  we won!!! love the land!!! #allin #cavs #champ...\n",
     "9        10      0   @user @user welcome here !  i'm   it's so #gr...\n",
     "10       11      0   Ã¢Â†Â •  #ireland consumer price index (mom) climb...\n",
     "11       12      0  we are so selfish. #orlando #standwithorlando ...\n",
     "12       13      0  i get to see my daddy today!!   #80days #getti...\n",
     "13       14      1  @user #cnn calls #michigan middle school 'buil...\n",
     "14       15      1  no comment!  in #australia   #opkillingbay #se...\n",
     "15       16      0  ouch...junior is angryÃ°ÂŸÂ˜Â •#got7 #junior #yugyo...\n",
     "16       17      0  i am thankful for having a paner. #thankful #p...\n",
     "17       18      1                               retweet if you agree! \n",
     "18       19      0  its #friday! Ã°ÂŸÂ˜Â€ smiles all around via ig use...\n",
     "19       20      0  as we all know, essential oils are not made of...\n",
     "20       21      0  #euro2016 people blaming ha for conceded goal ...\n",
     "21       22      0  sad little dude..   #badday #coneofshame #cats...\n",
     "22       23      0  product of the day: happy man #wine tool  who'...\n",
     "23       24      1    @user @user lumpy says i am a . prove it lumpy.\n",
     "24       25      0   @user #tgif   #ff to my #gamedev #indiedev #i...\n",
     "25       26      0  beautiful sign by vendor 80 for $45.00!! #upsi...\n",
     "26       27      0   @user all #smiles when #media is   !! Ã°ÂŸÂ˜ÂœÃ°ÂŸÂ˜...\n",
     "27       28      0  we had a great panel on the mediatization of t...\n",
     "28       29      0        happy father's day @user Ã°ÂŸÂ’Â“Ã°ÂŸÂ’Â“Ã°ÂŸÂ’Â“Ã°ÂŸÂ’Â“  \n",
     "29       30      0  50 people went to nightclub to have a good nig...\n",
```

```
    "...     ...    ...                                    ...\n",
    "31932  31933     0                          @user thanks gemma  \n",
    "31933  31934     1  @user judd is a  &amp;amp; #homophobic #freemilo #...\n",
    "31934  31935     1  lady banned from kentucky mall. @user  #jcpenn...\n",
    "31935  31936     0  ugh i'm trying to enjoy my happy hour drink &a...\n",
    "31936  31937     0  want to know how to live a   life? do more thi...\n",
    "31937  31938     0                            love island Ã°ÂŸÂ’Â”  \n",
    "31938  31939     0  my fav actor #vijaysethupathi ! my fav actress...\n",
    "31939  31940     0     whew  Ã°ÂŸÂ˜\n",
    " it's a productive and   #friday!!!\n",
    "31940  31941     0               @user she's finally here! @user   \n",
    "31941  31942     0  passed first year of uni #yay #love #pass #uni...\n",
    "31942  31943     0  this week is flying by   #humpday - #wednesday...\n",
    "31943  31944     0   @user modeling photoshoot this friday yay #mo...\n",
    "31944  31945     0  you're surrounded by people who love you (even...\n",
    "31945  31946     0  feel like... Ã°ÂŸÂ˜Â□Ã°ÂŸÂ□Â¶Ã°ÂŸÂ˜ÂŽ #dog #summer #hot #h...\n",
    "31946  31947     1  @user omfg i'm offended! i'm a  mailbox and i'...\n",
    "31947  31948     1  @user @user you don't have the balls to hashta...\n",
    "31948  31949     1   makes you ask yourself, who am i? then am i a...\n",
    "31949  31950     0  hear one of my new songs! don't go - katie ell...\n",
    "31950  31951     0   @user you can try to 'tail' us to stop, 'butt...\n",
    "31951  31952     0  i've just posted a new blog: #secondlife #lone...\n",
    "31952  31953     0               @user you went too far with @user  \n",
    "31953  31954     0  good morning #instagram #shower #water #berlin...\n",
    "31954  31955     0  #holiday   bull up: you will dominate your bul...\n",
    "31955  31956     0  less than 2 weeks Ã°ÂŸÂ˜\n",
    "Ã°ÂŸÂ™Â□Ã°ÂŸÂ□Â¼Ã°ÂŸÂ□Â¹Ã°ÂŸÂ˜ÂŽÃ°ÂŸÂŽÂµ @us...\n",
    "31956  31957     0  off fishing tomorrow @user carnt wait first ti...\n",
    "31957  31958     0  ate @user isz that
youuu?Ã°ÂŸÂ˜Â□Ã°ÂŸÂ˜Â□Ã°ÂŸÂ˜Â□Ã°ÂŸÂ˜Â□Ã°ÂŸÂ˜Â□Ã°...\n",
    "31958  31959     0    to see nina turner on the airwaves trying to...\n",
    "31959  31960     0  listening to sad songs on a monday morning otw...\n",
    "31960  31961     1  @user #sikh #temple vandalised in in #calgary,...\n",
    "31961  31962     0               thank you @user for you follow  \n",
    "\n",
    "[31962 rows x 3 columns]>"
   ]
  },
  "execution_count": 9,
  "metadata": {},
```

```
   "output_type": "execute_result"
  }
 ],
 "source": [
  "train.head"
 ]
},
{
 "cell_type": "code",
 "execution_count": 10,
 "metadata": {},
 "outputs": [],
 "source": [
  "combi = train.append(test,ignore_index = True,sort = True)"
 ]
},
```



CONCLUSION

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. Apache Spark proved prolific in extracting live streams of data and has further capability to store batches of data in HDFS and other major conventional storages. The processing capabilities of Spark makes the project flexible to further extend to multiple nodes, thereby supporting distributed computing. Real time data analysis makes it possible for business organizations to keep track of their services and generates opportunities to promote, advertise and improve from time to time.

REFERENCES

1. Efthymios Kouloumpis and Johanna Moore, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012 .

2. S. Batra and D. Rao, “Entity Based Sentiment Analysis on Twitter”, Stanford University,2010

3. Saif M.Mohammad and Xiaodan zhu ,Sentiment Analysis on of social media texts:,2014

4. Ekaterina kochmar, University of Cambridge, at the Cambridge coding Academy Data Science.2016

5. Manju Venugopalan and Deepa Gupta ,Exploring Sentiment Analysis on Twitter Data, IEEE 2015