Eindhoven University of Technology

2IMW15
Web information retrieval and Data Mining

Guliz Cosan Mutlu - 1034515

1. General Approach

As rumors will be investigated based on topics of tweets, it is necessary to first find out tweets that are related to same topic. Discovery of each tweet's topic is essential since further rumor analysis will be performed by analyzing all tweets that has same topic. For this purpose, topic classification, which is the most important part of the overall project, is performed using different approaches. To classify topics, the tasks that are explained in detail later in this document are implemented.

Before starting to process data, a set of operations are done for text normalization.

- Only English words are considered.
- Only certain characters (ASCII letters, digits etc.) are accepted.
- Single words that are smaller than 3 words are removed.
- Stemming is applied by using nltk.stem library (for supervised learning only).

2. TF-IDF

2.1. Motivation

In order to work with text data for topic classification, a transformation to numerical vectors should be done. This is achieved by using TF-IDF.

2.2. Description and Goal

TF-IDF is a method of downscaling. This combined statistic has two different components that are multiplied. These components are namely TF (Term Frequency) and IDF (Inverse Document Frequency). This method is useful to eliminate over-emphasized words. TF is a simple measurement of the frequency of a word in a document. As relative frequency would give more insightful results as many documents will be examined, absolute frequency of occurrence of a specific word is divided by the total number of all words. A vector is generated by applying this method to all words in a document. Yet, it is necessary to multiply TF vector with another measure called IDF to eliminate the misleading TF vector of words that are used too frequently among all the documents such as "an", "of", "the", "but" etc. The idea behind IDF (Inverse Document Frequency) is to raise the importance of less frequent words and to decrease

the importance of too frequent words. Similarly, IDF is built as a vector where each component of this vector illustrates the measure of occurrence of each word (within specific input text) among all the documents [1].

Based on the results of this calculation, I infer the most frequent words in each document and cluster documents that has similar topics.

2.3. Implementation

Above goal is achieved by using sklearn and nltk libraries. Sklearn library has TfidfVectorizer utility which calculates TF-IDF vectors.

2.4. Results

This task enables us to see most frequent words in each topic cluster. TF-IDF calculations are not shown as values but instead, the program helps the classification approach to show top 10 most frequent words in the document. For example, for Multinomial Naïve Bayes, the most significant words are calculated as below.

```
Most significant words per category

alt.atheism: would islam religion atheist moral think peopl one say god
comp.graphics: format anyon know look use program thank imag file graphic
comp.os.ms-windows.misc: know run program thank problem do driver use file window
comp.sys.ibm.pc.hardware: system disk control thank ide scsi bu use card drive
comp.sys.mac.hardware: work one get thank simm use problem drive appl mac
comp.windows.x: program display applic run thank widget motif server use window
misc.forsale: email price pleas condit new includ sell ship offer sale
rec.autos: good look drive dealer one engin would get like car
rec.motorcycles: know rider helmet get like dod one motorcycl ride bike
rec.sport.baseball: think win run hit player basebal pitch year team game
rec.sport.hockey: year win nhl season playoff player hockey play team game
sci.crypt: escrow nsa would govern use secur chip clipper encrypt key
sci.electronics: work get know like power anyon one would circuit use
sci.med: test rico puerto report death case first mosquito viru zika
sci.space: get shuttl one like moon would launch nasa orbit space
soc.religion.christian: christ bibl peopl one believ would church jesu christian god
talk.politics.guns: fbi get right fire law firearm peopl weapon would gun
talk.politics.mideast: would one turkish kill peopl armenian jew arab isra israel
talk.politics.misc: think clinton make homosexu one state govern would tax peopl
talk.religion.misc: believ bibl think one would peopl say jesu god christian
```

Figure 1: Most Significant Words per Category for 20NewsGroup and ZikaVirus datasets

3. Topic Classifier

As I have two different data, one of them is labelled and the other is not, I performed both supervised and unsupervised learning approaches to implement topic classification task. These two learning methods are explained in detail below:

3.1 Topic Classification with Supervised Learning Approach

3.1.1. Motivation

The ultimate motivation of this task is to predict topic of an input tweet per predetermined topics which are the labels of 20 newsgroups data.

3.1.2. Approach

The classification is implemented with 5 different supervised learning methods that are available in sklearn library of Python. These methods are:

- Multinomial Naïve Bayes Classifier[2]
- Bernoulli Naïve Bayes Classifier[2]
- Stochastic Gradient Descent (SGD) Classifier
- Linear Support Vector Classifier (SVC)
- Ridge Classifier

The performances of above classifiers are compared and it is concluded that Multinomial Naïve Bayes Classifier is found to be the most accurate. Basically, this approach calculates the probability distribution of input tweet belonging to each of predetermined topics and then assigns the tweet to the topic that has the highest probability.

3.1.3. Data Exploration

In this part, supervised learning algorithms are used as mentioned before. So, in order to

obtain more accurate results and achieve the goal, a well labeled data is needed. At first the data that is provided with name ‘dataset_2016’ (4200 tweets about Zika Virus in tweets.csv was processed) is considered but it was insufficient individually since it only focuses on one label. Therefore, I have also used ’20 newsgroups’ dataset which is freely available on internet[3]. It is a collection of about 20,000 newsgroup documents with the following categories:

- alt.atheism
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- misc.forsale
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc
- talk.religion.misc

As the names of these categories clearly indicate the context itself, I decided to use same names for the topic labels. The data that is provided is related to Zika Virus, therefore its content is also labelled to ‘sci.med’ which is the label related to health.

3.1.4. Implementation

For the implementation of supervised learning algorithms; nltk, pickle and sklearn libraries are utilized. As two different data sources are used, they are merged as the first step. The merged data is shuffled and trained by using 30% of the whole data as test data and the rest as training data.

As mentioned before, 5 different classifiers are implemented and tested. The performances are shown in Results & Evaluation part.

3.1.5. Results & Evaluation

The performance results of 5 different classifiers are shown below:

Classifier	Accuracy
Multinomial Naïve Bayes Classifier	0.7063
Bernoulli Naïve Bayes Classifier	0.5636
Stochastic Gradient Descent (SGD) Classifier	0.7046
Linear Support Vector Classifier (SVC)	0.6946
Ridge Classifier	0.7014

As stated before, Multinomial Naïve Bayes Classifier has the best performance. Here, results of Multinomial Naïve Bayes Classifier will be evaluated in detail. In order to show its performance, a confusion matrix can be generated as a visual illustration and also as number matrix by using the program.

Below, the visual illustration and the number version of confusion matrix is shown respectively:

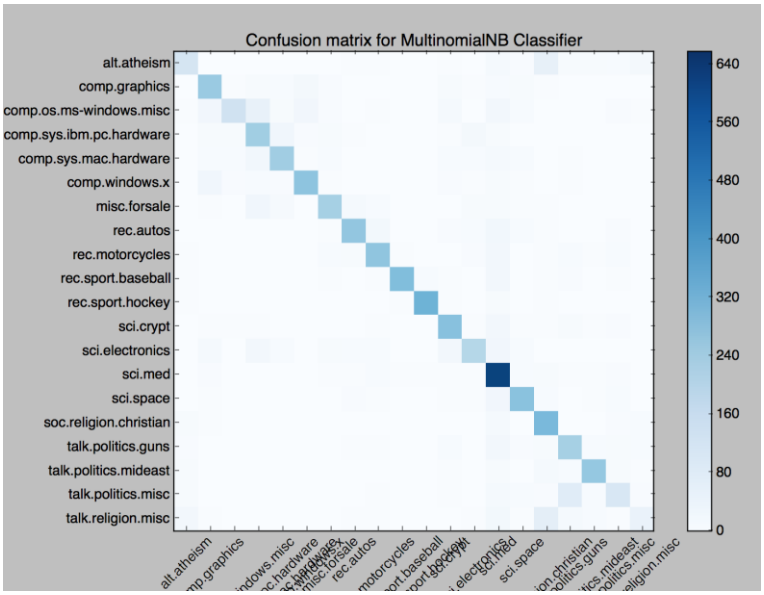

Figure 2: Visual Illustration of Confusion Matrix based on results of Multinomial Naïve Bayes Classifier

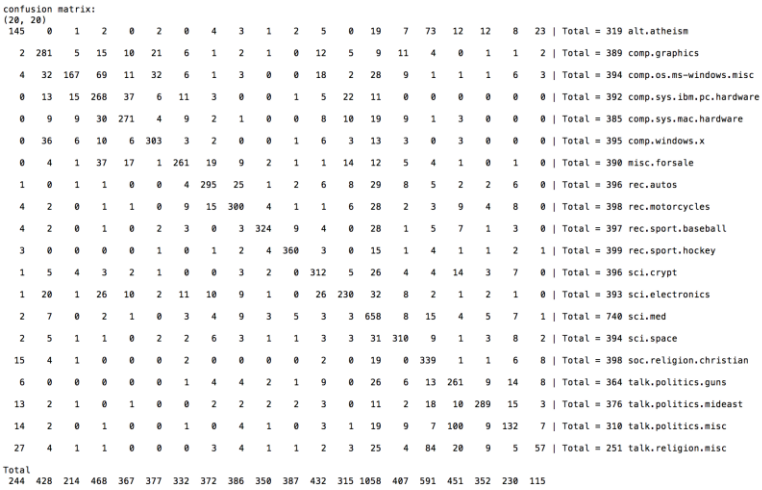

Figure 3: Confusion Matrix based on results of Multinomial Naïve Bayes Classifier

3.2 Topic Classification with Unsupervised Learning Approach

The data in hand, which is around 20.42 GB, is crawled from Twitter already on first quartile. As the data is too large for processing, a preprocessing is applied. The data is then used for topic extraction as an option. As the general task was related to topic classification from the beginning, this part is also implemented.

3.2.1. Motivation

The main aim is to extract topic clusters from unlabeled data via unsupervised learning techniques.

3.2.2. Approach

In order to extract topics of the corpus of documents (in our case they are tweets), Non-negative Matrix Factorization and Latent Dirichlet Allocation are applied.

These methods are:

- Latent Dirichlet Allocation (LDA): This algorithm assumes that each document includes

a set of different topics and groups the documents accordingly[4].

- Non-negative Matrix Factorization (NMF): This algorithm clusters documents that has same features by data dimension reduction.

3.2.3 Implementation

These two algorithms are implemented by using sklearn library of Python. The output is a list of topics, each represented as a list of terms.

3.2.3. Results

The unlabeled data is grouped into 20 topics and most important 10 words of the topics are shown as output. As the data does not have an already clustered version, it is unfortunately not possible to test the performance of the results.

```
Topics in LDA model:
Topic #0:
gon na help follow girl believe things police little makes school face early biggest girls open goal cleveland turn worth
Topic #1:
tell job bad proud end debate poll friends physics stay chemistry forget final yesterday candidate comes safe corruption stand awarded
Topic #2:
big media support family state breaking coming wants run weekend future words perfect leave act huge killed security criminal dress
Topic #3:
justin campaign years party wait heart email chicago wikileaks supporters guy seen ago bieber ahead reason damn dressed sick thinking
Topic #4:
god yes house woman hey government change mean wrong special voted listen law hrc literally ask sad report ta public
Topic #5:
news emails week morning tweets high votes thousand waiting court bernie movie crazy mom list record daily justice visit tells
Topic #6:
fbi come thing old american states comey lost americans finally russia investigation share child king wanted machines wife huma past
Topic #7:
better wan song na money fan fight case corrupt hours office melania college link article fake celebrate clear shawn online
Topic #8:
live retweet set shows lose ass goes half plan sex trust death belieber crowd months usa worst park bit yall
Topic #9:
real team november hard play true omg needs told single photo beat saw forever nation human deserves welcome nov speech
Topic #10:
tweet said national miss country sure remember china costume 1st watching fuck history 100 gaga foundation chance congratulations playing enjoy
Topic #11:
stop women feel start free away matter talk fun times hate dont left making soon excited million winning far congrats
Topic #12:
baby read amazing looks use looking shit talking went cute fucking break join latest awesome hashtag super line star cnn
Topic #13:
president twitter ready north person trying actually voters rally lot bring florida book using knows feeling cool costumes idea calls
Topic #14:
election music home favorite long rock thought album hit working fact children came female anti peace count giving imagine radio
Topic #15:
beliebers snapchat story check season kids friend saying point hear war used head attack purpose deal maine selena pls friday
Topic #16:
days black tomorrow power lead wow city 000 gets means brexit ozil close pay warriors running took takes tweeting voter
Topic #17:
guys white cubs post wins cause sorry nice lady moment question lets update red respect isis dear agree blue saudi
Topic #18:
reef care dead yeah men taking late hell seeing democrats pic living heard phone queen lives instead hands messi add
Topic #19:
fans getting thanksgiving beautiful wish gop place try smile series podesta united mind boy polls john important truth strong event
```

Figure 4: Extracted Topics in LDA Mode

```
Topics in NMF model:
Topic #0:
happy day hope birthday miss bday today girl ily wish beautiful best lots luv thanks pretty deserve hbd friend soon
Topic #1:
year artist voting 2016 collaboration ariana valid tour voto rts old retweet arianators votes 100 selena ama video best queen
Topic #2:
halloween happy costume costumes candy party night dressed best fun dress 2016 year spooky safe kids little weekend treat trick
Topic #3:
love life song heart baby forever girl guys true justin thanks beautiful lots miss best really way people fall happiness
Topic #4:
trump donald president election poll debate women says clinton america supporters rally campaign said media melania people gop say pence
Topic #5:
nobel prize physics chemistry 2016 awarded machines win matter scientists peace molecular trio world british exotic work wins smallest goes
Topic #6:
na gon wan say lol tonight think stop really cause tell miss alright leave win watch right live talk guys
Topic #7:
vote retweet artist valid beliebers ema tweet forget 2016 counts count need guys ariana mtv justin hillary day early hashtag
Topic #8:
like looks look feel lol people got really sounds think right feels going said barrier man looking waiting life act
Topic #9:
time long hope today thanks best right think got spend game say tell come watch tonight lol barrier remember night
Topic #10:
hillary clinton fbi campaign emails obama email wikileaks president foundation breaking says investigation poll debate people election state comey crooked
Topic #11:
know need people right lol think really dont things going let got say better thing real means life cause best
Topic #12:
thank god support amazing guys night today day making sharing sir best omg wow week lord job bless appreciate friend
Topic #13:
good morning luck night day job look today really work looks news god people thing man guys looking bad thanks
Topic #14:
favorite rock pop female artist soul male music radio home voting album song retweet country selena 2016 voto purpose duo
Topic #15:
christmas thanksgiving november music ready 1st excited merry wait days start season tree movies means holiday officially basically listening early
Topic #16:
let win justin beliebers voting guys power retweet bieber follow tweet game going best tonight need today come proud help
Topic #17:
new video york watch album poll check music song week today brand single tonight times post peak episode city selena
Topic #18:
make america sure people better feel going proud happen president right things think need god life way smile world let
Topic #19:
want really people win follow retweet say life president dont hear justin watch live bad come tell talk world hug
```

Figure 5: Extracted Topics in NMF Model

Bibliography

1. Ramos, J. (n.d.). *Using TF-IDF to Determine Word Relevance in Document Queries.* Piscataway, NJ: Department of Computer Science, Rutgers University.

2. *scikit-learn.* (2010). Retrieved 01 11, 2017, from http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

3. *Naive Bayes text classification.* (2008). Cambridge University Press.

4. David M. Blei, A. Y. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.