Significant News Detection

Master's Thesis

Diwas Sharma

October 26, 2018

The University of Alabama in Huntsville

Table of contents

Introduction

Motivation

Based on a recent study,[1]

- 68 percent of US adults at least occasionally get news on social media.
- However, 57 percent of those people expect the news to be largely inaccurate.

[1] Katerina Eva Matsa and Elisa Shearer. "News Use Across Social Medial Platforms 2018". In: Pew Research Center, Journalism and Media (2018).

False news

News examples obtained from The Onion,

1. "**104-Year-Old Reveals Secret To Long Life Being Cursed By Witch To Wander Earth Eternally**"
2. "**Study Finds Over 5 Million Birds Die Annually From Head-On Collisions With Clouds**"

What is fake news

"Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design."[2]

[2] Axel Gelfert. "Fake news: a definition". In: Informal Logic 38.1 (2018), pp. 84–117.

Fake news

Fake news viewing impacts

- political attitudes toward politicians[3]
- the ability of people to accept truthful news [4].

[3] Meital Balmas. "When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism". In: Communication Research 41.3 (2014), pp. 430–454.

[4] https://www.nytimes.com/2016/11/28/opinion/fake-news-and-the-internet-shell-game.html?%20r=0

Fake news detection

It is impractical for human reviewers to manually categorize every single article, statement or message.

Automatic fake news detection

Challenges,[5]

- Difficult to classify solely on content
- Auxiliary information might be unreliable

[5] Kai Shu et al. "Fake news detection on social media: A data mining perspective". In: ACM SIGKDD Explorations Newsletter 19.1 (2017), pp. 22–36.

Important news

Articles on topics such as **infotainment, personal news, health and beauty tips, etc** are somewhat less important so might not matter even if they were fake.

For example consider following sentences,

- "I watched the Black Panther movie in the weekend. Fantastic effects!".
- "There has been a shooting at the mall"

Hard/Soft news

- **Hard news:** reports about politics, science, technology and related topics.[6]
- **Soft news:** reports about celebrities, sport and other entertainment-centred stories.[7]
- Inadequate for the purpose of identifying important article from unimportant ones.

[6] Carsten Reinemann et al. "Hard and soft news: A review of concepts, operationalizations and key findings". In: Journalism 13.2 (2012), pp. 221–239.

[7] Ibid.

Significant news

A text is labelled as significant if it

- affects a large number of people
- changes the routines of daily life
- needs verification on the information presented

Research problems

1. Studying existing fake news detection methods for detecting significant news.
2. Building a new dataset for significant news detection.
3. Selecting features from fake news detection methods and evaluating a number of classifier models for significant news detection.

Related works on fake news classification

- **Paper:** Fake news detection using naive Bayes classifier
  Classification: Binary
  Model: Naive Bayes
  Accuracy: 74%
- **Paper:** Evaluating machine learning algorithms for fake news detection
  Classification: Binary
  Models: Random Forest, Bounded decision tree, Gradient boosting, SVM, SGD
  Best model: SGD classifier
  Best accuracy: 77.2%

Related works on fake news classification

- **Paper:** "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection
  Classification: Multiclass
  Labels: "pants-fire", "false", "barely-true", "half-true", "mostly-true", "true"
  Models: Logistic regression, SVM, Bidirection LSTM, CNN
  Best model: CNN
  Best accuracy: 27.0%

Related works on fake news classification

- **Paper:** Neural User Response Generator: Fake News Detection with Collective User Intelligence.
  Classification: Binary
  Models: SVM, CNN, TCNN, TCNN-URG
  Best model: TCNN-URG
  Best accuracy: 89.94%

Method

Dataset

- Pre existing dataset is not available
- Need to create a new dataset
- Source: official social account of police departments
  - will contain a large number of significant news articles.
  - the content will be truthful and verified.

Data Acquisition

Twitter Account	Country
NYPD News	USA
Metropolitan Police	UK
Victoria Police	Australia
Seattle Police Dept	USA
NYPD Counterterrorism	USA

Table 1: Twitter accounts used for collecting data

Dataset labelling

Each article was then manually labeled as either significant news or non-significant news based on the pre-conditions.

- Does it affect large number of people ?
- Does it change the routines of daily life ?
- Do we need to verify it ?

Labelling sample 1

Text: *Detectives investigating the murder of Kwabena Nelson in Tottenham have made an arrest Haringey.*

Label: Significant

Reason:

affects a large number of people	✓
changes the routines of daily life	✓
needs verification	✓

Labelling sample 2

Text: *Officers investigating shooting in 8100 blk 31st Ave SW. Adult male victim taken to HMC with serious injuries. Update soon.*

Label: Significant

Reason:

affects a large number of people	✓
changes the routines of daily life	✓
needs verification	✓

Labelling sample 3

Text: *Great example of NYPDconnecting in the Bronx. NYPD49Pct NeighborhoodPolicing officers worked with the community to address a garbage condition on a resident block in their neighborhood.*

Label: Non-Significant

Reason:

affects a large number of people	?
changes the routines of daily life	×
needs verification	✓

Labelling sample 4

Text: *It was over before it began for this 18-year-old who lost her licence after only two hours.*

Label: Non-Significant

Reason:

affects a large number of people	×
changes the routines of daily life	×
needs verification	✓

Dataset

Label	Number of samples
Significant	1548
Non-significant	595
	2143

Table 2: Significant news dataset statistics

Feature generation

Steps

1. Tokenization
2. Stopword filtering
3. Stemming
4. TF-IDF

Tokenization

Text:

Watch: @PIX11News gives an inside look at our #NeighborhoodPolicing meetings on how they are connecting local NYPD police officers with the community. `https://t.co/D6K5DWxWWm` `https://t.co/NM0AWpPgja`

Tokens:

"watch", "pix", "news", "gives", "an", "inside", "look", "at", "our", "neighborhoodpolicing", "meetings", "on", "how", "they", "are", "connecting", "local", "nypd", "police", "officers", "with", "the", "community"

Stop word filtering

Stop words:

"i", "me", "the", "sunday", "monday", "January", "Februray", "nypd", "nypdct", etc

Filtered Tokens:

"watch", "pix", "news", "gives", ~~"an"~~, "inside", "look", ~~"at"~~, ~~"our"~~, ~~"neighborhoodpolicing"~~, "meetings", ~~"on"~~, ~~"how"~~, ~~"they"~~, ~~"are"~~, "connecting", "local", ~~"nypd"~~, "police", "officers", ~~"with"~~, ~~"the"~~, "community"

Stemming

Porter's[8] algorithm was used to reduce the inflected word to their stems.

Stemmed Tokens:

"watch", "pix", "news", "gives", ~~"an"~~, "inside", "look", ~~"at"~~, ~~"our"~~, ~~"neighborhoodpolicing"~~, "meetings", ~~"on"~~, ~~"how"~~, ~~"they"~~, ~~"are"~~, "connecting", "local", ~~"nypd"~~, "police", "officers", ~~"with"~~, ~~"the"~~, "communitiy"

[8] Martin F Porter. "An algorithm for suffix stripping". In: Program 14.3 (1980), pp. 130–137.

TF-IDF

TF-IDF[9] was used to obtain the feature vector for the text

$$tfidf(t, d) = tf(t, d) * idf(t) \tag{1}$$

where,
$tf(t, d)$ is the count of the term in the document
$idf(t)$ is the measure of information that the term provides.

[9] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: Journal of documentation 28.1 (1972), pp. 11–21.

IDF

The *idf* for a term is calculated as:

$$idf(t) = \log \frac{N}{1 + df(t)} \tag{2}$$

where,

N is the number of total number of documents

$df(t)$ is the number of documents that contains the term t.

The IDF values for all the stems are calculated and the stems are sorted in descending order of IDF values.

IDF values

stem	IDF value	stem	IDF value	stem	IDF value
graffiti	7.46	ring	7.33	harrow	7.33
colour	7.33	mother	7.21	food	7.21
tragedi	7.21	veteran	7.21	bay	7.21
leg	7.21	tuozzolo	7.21	church	7.21
design	7.21	termin	7.21	vote	7.21
mornington	7.21	psos	7.21	brief	7.21
babi	7.1	race	7.1	pull	7.1
trade	7.1	post	7.1	coordin	7.1
class	7.1	extra	7.1	complaint	7.1
ps	7.1	came	7.1	worth	7.1
teamwork	7.1	adult	7.1	civilian	7.1
notic	7.1	collaps	7.1	brown	7.1
greatest	7.1	terenc	7.1	count	7.1
medic	7.1	plenti	7.1	akay	7.1
one	7.1	citizen	7.1	noth	7.1
strand	7.1	page	7.1	alcohol	7.1
true	7.1	thief	7.1		

Figure 1: Top 50 IDF values

TF-IDF vector

- The top 1000 stems are selected from the sorted stem list.
- Can convert a text to a 1000 dimensional vector.
- Each component of the vector is a TF-IDF value for a term t in the document d.
- The vector v obtained is then normalized using the L^2 norm as follows,

$$v_{norm} = \frac{v}{\|v\|} \tag{3}$$

Classification models

1. Logistic Regression
2. Support Vector Machine
3. Random Forests
4. Neural Network

Logistic Regression

- Hypothesis function

$$h_\theta(x) = \frac{1}{1 + e^{-\theta t . x}} \tag{4}$$

- Regularized logistic regression with L2 penalty
- The best value of the regularization coefficient λ is selected during the cross-validation
- λ values: 1, 50, 100, 200, 300, 400

Support Vector Machine

- Soft margin
- RBF kernel

$$K(x, x') = exp(-\gamma \|x - x'\|^2) \tag{5}$$

- γ: 0.001
- The best value of the hyperparameter C is selected during the cross-validation
- C values: 1, 50, 100, 200, 300, 400

Random Forests

- Split criterion: Gini
- Minimum sample split: 2
- Minimum sample leaf: 1
- The best value for the number of trees is selected during the cross-validation step
- Number of trees: 10, 12, 24

Neural network

- Three layer network
- Activation function: ReLu
- Optimizer: ADAM
- Mini batch size: 200
- The best value for the learning rate and number of hidden units will be selected during the cross-validation step.
- Learning rates: 0.1, 0.01
- Hidden units: 250, 500

Training and testing

- The dataset is seperated into training set and testing set as follows,
  - **Training:** 1714 samples (80%)
  - **Testing:** 429 samples (20%)
- For each classification model, a random set of hyperparameters is selected.
- The performance of the model with the hyperparameters is evaluated using the 5 fold cross validation
- The set of hyperparameters that obtained the best validation performance is selected.

Training and testing

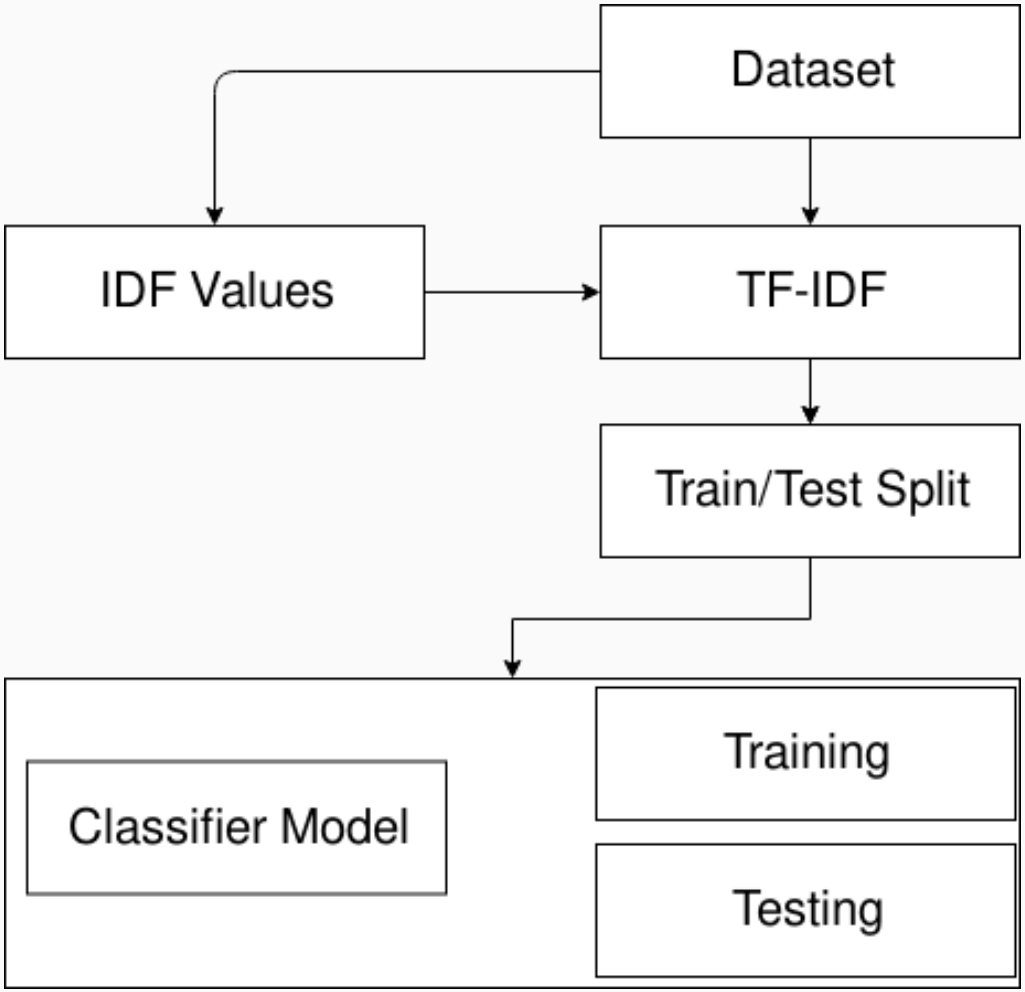

Figure 2: Flow diagram for training and testing

Prediction

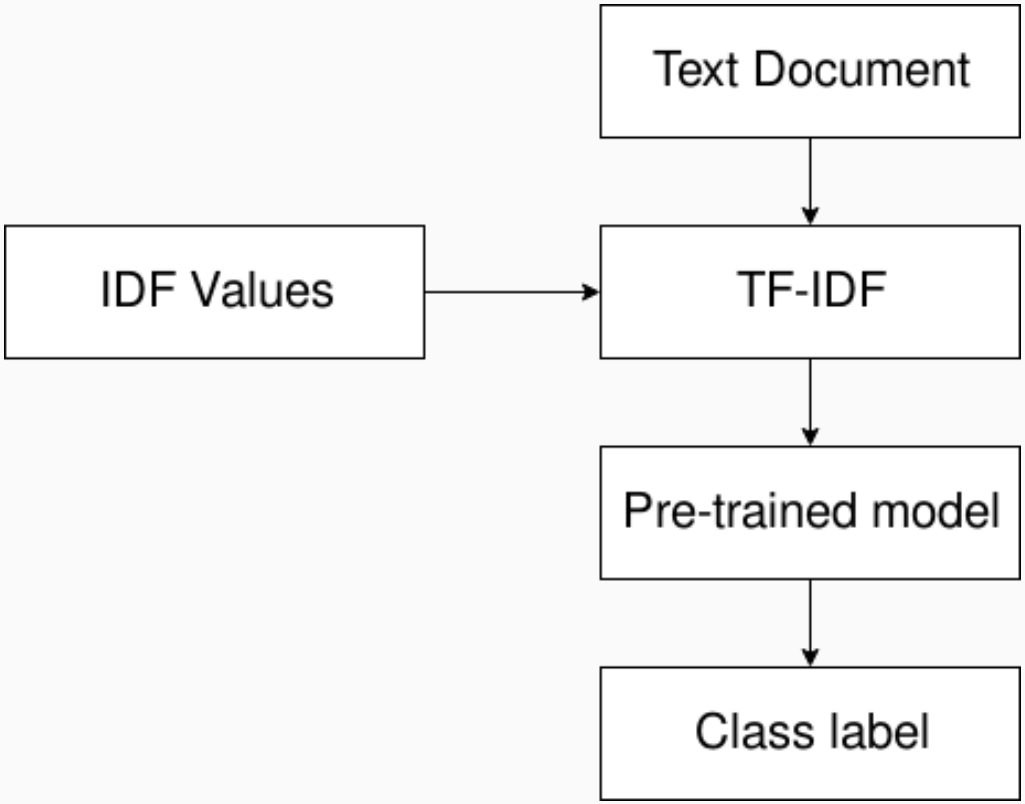

Figure 3: Flow diagram for prediction

Results

Dataset visualization

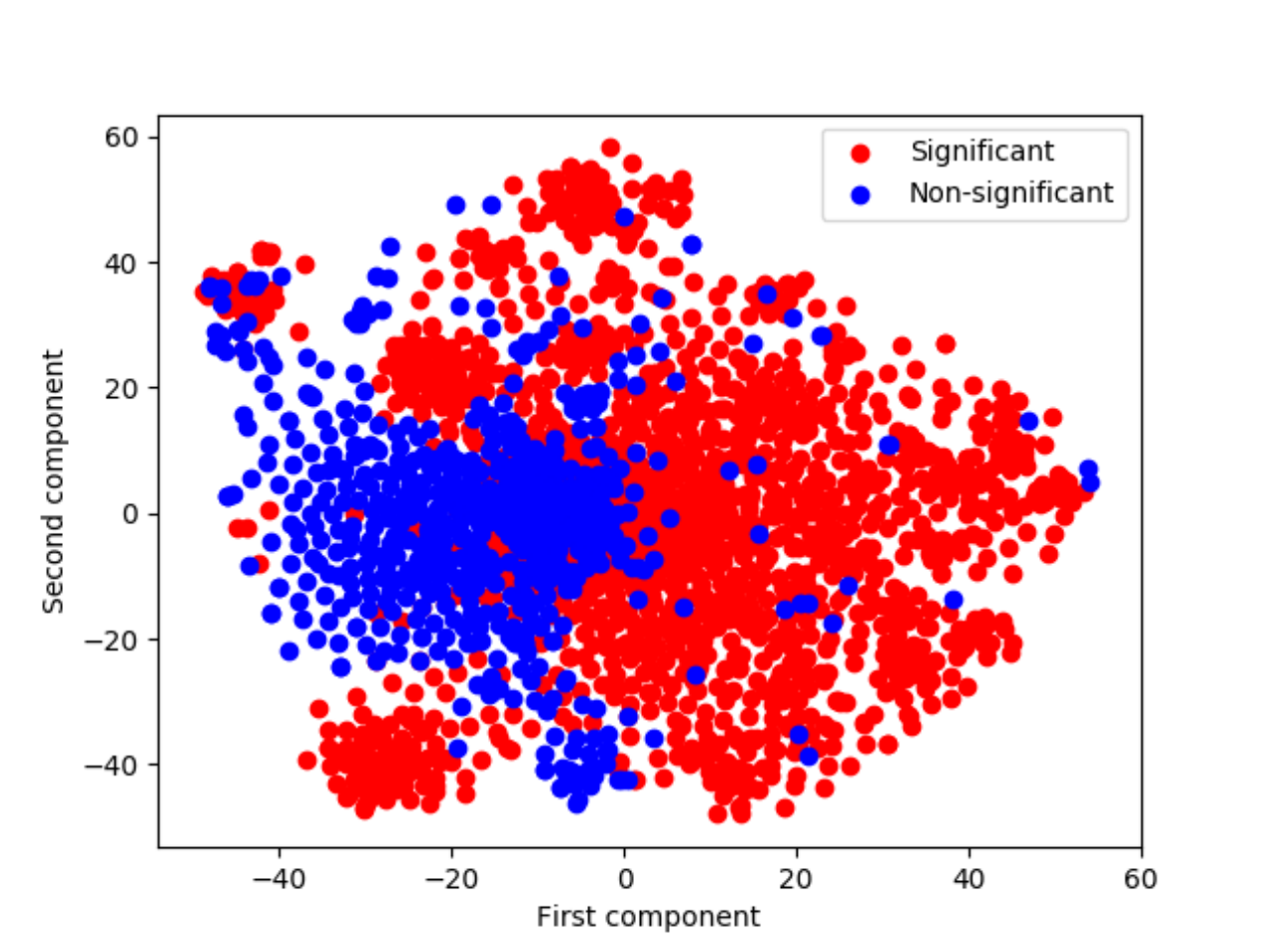

Figure 4: Dataset visualization using t-SNE

Test performance

Model	Accuracy	Precison	Recall	Fscore
Regularized Logistic Regression	92.81	91.224	90.02	90.852
Random Forests	92.022	90.36	90.02	90.128
SVM	93.448	92.468	91.464	91.128
Neural Network	**93.654**	91.602	91.685	**91.43**

Figure 5: Average performance of classifiers on significant news dataset

Correct Predictions

Actual: Significant **Predicted: Significant**

1. Detectives are investigating following a robbery outside a licensed premises in St Albans overnight.More..
2. Homicide Squad detectives have charged a man following an alleged fatal stabbing at Ultima.
3. Detectives investigating bank robbery approx. 2:15 pm in 600 blk S. Dearborn. Suspect fled, at large. He is described as white male, 30s, 5'7", dark bb cap, grey hooded sweatshirt. If you recognize him, contact SPD. https://t.co/Fg4Cblhhxd.

Correct Predictions

Actual: Non-significant **Predicted: Non-significant**

1. RT seabikeblog: Had a bike stolen recently? May want to SeattlePD's GetYourBikeBack timeline. Just posted a bunch. SEAbikes
2. There's two weeks left to vote in the Met Excellence Awards. MPSTootingTnC are nominated for Safer Neighbourhoods Team of the Year for their diligent work tackling an increasing drug anti-social behaviour issue in the area. Learn more; cast your vote
3. It was over before it began for this 18-year-old who lost her licence after only two hours.

Incorrect Predictions

Actual: Non-significant **Predicted: Significant**

1. UK head of counter terrorism policing calls on everyone to play role in defeating terrorism and extremism https://t.co/rfVzB8jF76 https://t.co/KwzBAfHJ4B
2. RT @NYPDnews: "The assassination of my brother in 1988 was not only a personal tragedy, but was really a terrible tragedy for the entire NY...
3. Police Life magazine recently headed to Mornington Peninsula for an evening shift in the brawler van that covered all bases from bitumen to bay. Check out the full story in the Police Life Summer edition

Incorrect Predictions

Actual: Significant **Predicted: Non-significant**

1. RT @MPSCamden: Police were dealing with a suspicious package in Pentonville Road/Caledonian Road, KingsCross. The package has been deemed...
2. RT @MPSHammFul: FH schools officers carried out a weapons sweep in Little Wormwood Scrubs glad to say nothing found StopKnifeCrime OpScep...
3. All roads in Melbourne's CBD have been reopened after a large number of pedestrians were struck by a car on Flinders Street yesterday afternoon. All trams are now also operating as per usual.

Conclusion

Summary

- Defined significant news
- Prepared a new dataset based on the definition
- TF-IDF algorithm was used for feature generation
- Multiple classifiers were trained
- The classifiers were then evaluated on the test set
- The test performance of the classifiers was fairly good

Future works

- Refine the data further and add more samples
- Improve the definition for significant news
- Improving the feature extraction methods: Higher n-gram models, Word embedding models, etc
- Classification algorithms: HMM, GRU, LSTM, CNN etc

Questions?