Baidu Security

Tracking the Criminal of Fake News Based on a Unified Embedding

lvzhonghou@baidu.com wangwenhua@baidu.com
goodmandou@gmail.com liyiping02@baidu.com

Baidu Security

Contents

Baidu Security

Part-01
Introduction

Social Media Usage

- **3.8 million** of people use social media.
- Internet users worldwide spent an average of **144 minutes per day** on social media platforms in 2019.[1]

> *A lie will travel around the world while the truth is pulling on its boots.*
>
> —— Mark Twain

False Information

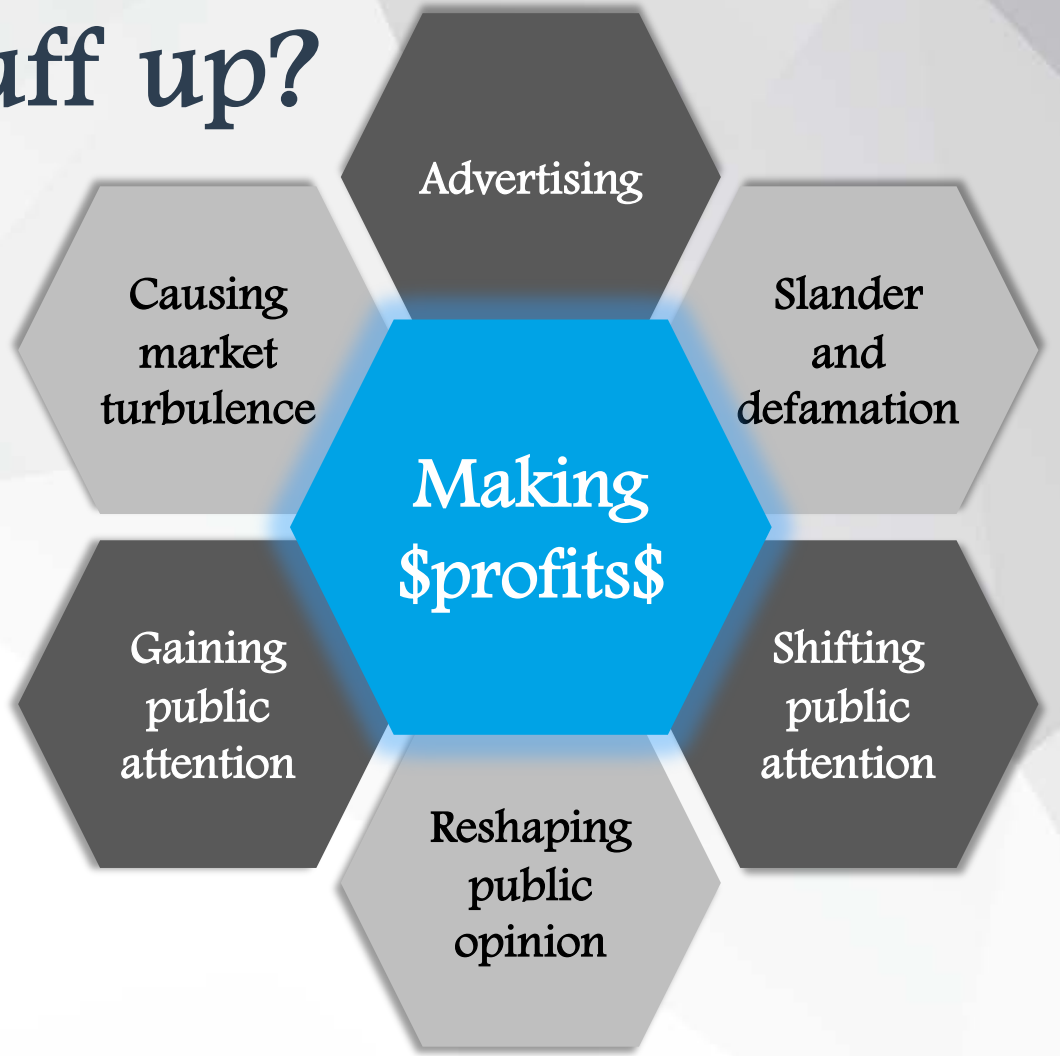
- **Spreads widely**: False news stories are 70% more likely to be retweeted on Twitter than true ones.[2]
- **Hard to detect**: People tend to believe what they want to believe, and fake news are intended to evoke strong emotional reactions.

Why people just make stuff up?

Public Opinion Attack is a PR move that alters public opinions by fabricating rumors and fake news.

Online Water Army is a group of Internet ghostwriters paid to post online comments with particular content.

Online Short Attack is a typical public opinion attack that anonymously posts fake news.

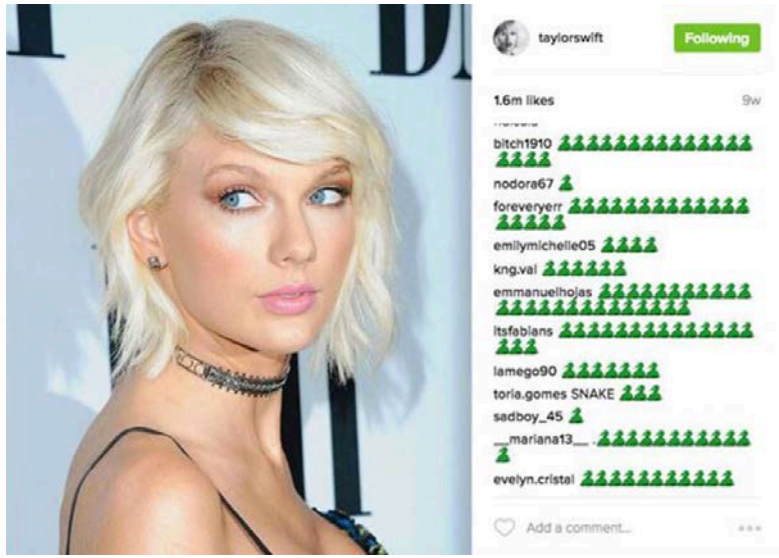


01 Introduction

Victims

Politicians

Companies



Celebrities

Consumers

Real world Consequences

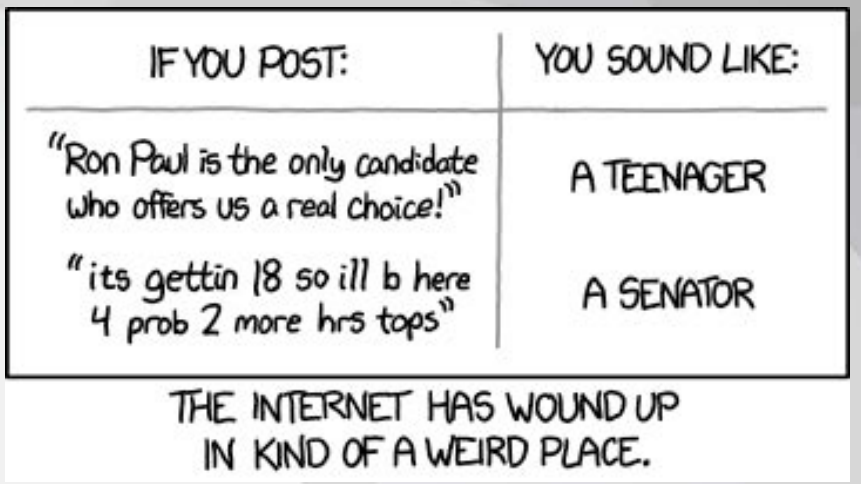
Tracking anonymous articles is …

Challenging

Necessary

01 Introduction

However, it is doable since each author has a **WRITING STYLE**…



Lexicon	01	Vocabulary Richness; Word Frequencies; Word N-grams
Character	02	Character Types; Character N-grams; Compression Methods
Syntax	03	Part-of-Speech; Sentence and Phrase Structure
Semantics	04	Synonyms; Semantic Dependencies; Functional

Baidu Security

Part-02
Related Works

02 Related Works: Authorship Analysis

Authorship Analysis contains three sub-problems:

Authorship Attribution

The article's author is identified from a given list of candidates.

Authorship Verification

Whether the authors of two articles are the same or not.

Authorship Clustering

Articles of the same author are clustered.

The COMBINATION of approaches to these three sub-problems is applied to track anonymous articles

02 Related Works: Authorship Analysis

In the early 2000, the method of stylistic and content features (lexical, character, syntactic, and semantic) extraction was proposed. [4] [5]

In recent years, method based on deep learning networks was proposed for AA. [9] [10] [11]

Initiated by *"Inference & Disputed Authorship"*, the work of Mosterller and Wallace.[3]

Around 2010, character n-gram was proven to be the most effective feature for AA. [6] [7] [8]

Baidu Security

Part-03
Method

03 Method: Overview

A unified method for authorship attribution, author verification, and author clustering

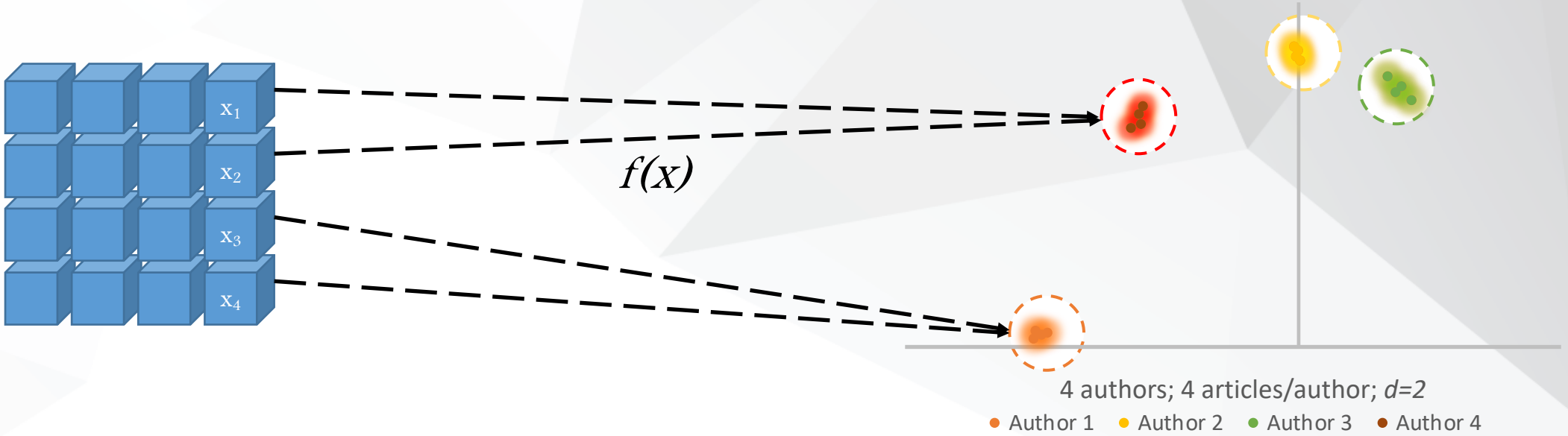
- Inspired by the FaceNet[12] (face recognition, face verification, face clustering)

- Based on deep learning network, to learn a mapping *f(x)* from texts to compact *n*-dimensional Euclidean space *R*

- Euclidean distances in the space indicate text similarity

03 Method: Overview

A unified method for authorship attribution, author verification, and author clustering

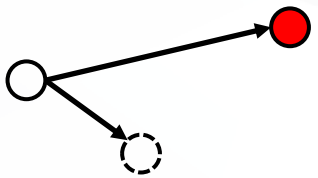


f(x) embeds a text *x* into the surface of a sphere with a radius of 1 and a center of origin in *R*

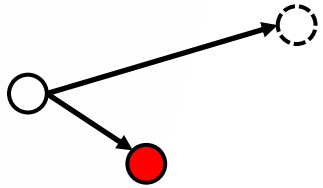
Texts of the same author have small distances and texts of the distinct author have large distances

03 Method: Loss Function

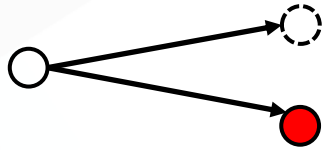
A triplet x_i is defined as $< x_i^a, x_i^p, x_i^n >$



$d_{ap} < d_{an}$



$d_{ap} > d_{an}$



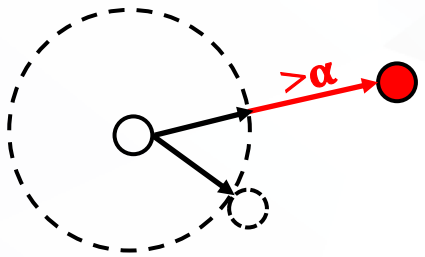
$d_{ap} = d_{an}$

- An anchor x_i^a (the target text)
- A positive x_i^p (a text of the same author as the anchor)
- A negative x_i^n (a text of an author different from the anchor)

$d_{ap} = || f(x_i^a) - f(x_i^p) ||_2^2$ $\quad$ $d_{an} = || f(x_i^a) - f(x_i^n) ||_2^2$

03 Method: Loss Function

Triplet Loss aims to separate the anchor and positive from the negative, which let ($d_{ap} + \alpha < d_{an}$)



$d_{ap} + \alpha < d_{an}$

$d_{ap} > d_{an}$

$d_{ap} = d_{an}$

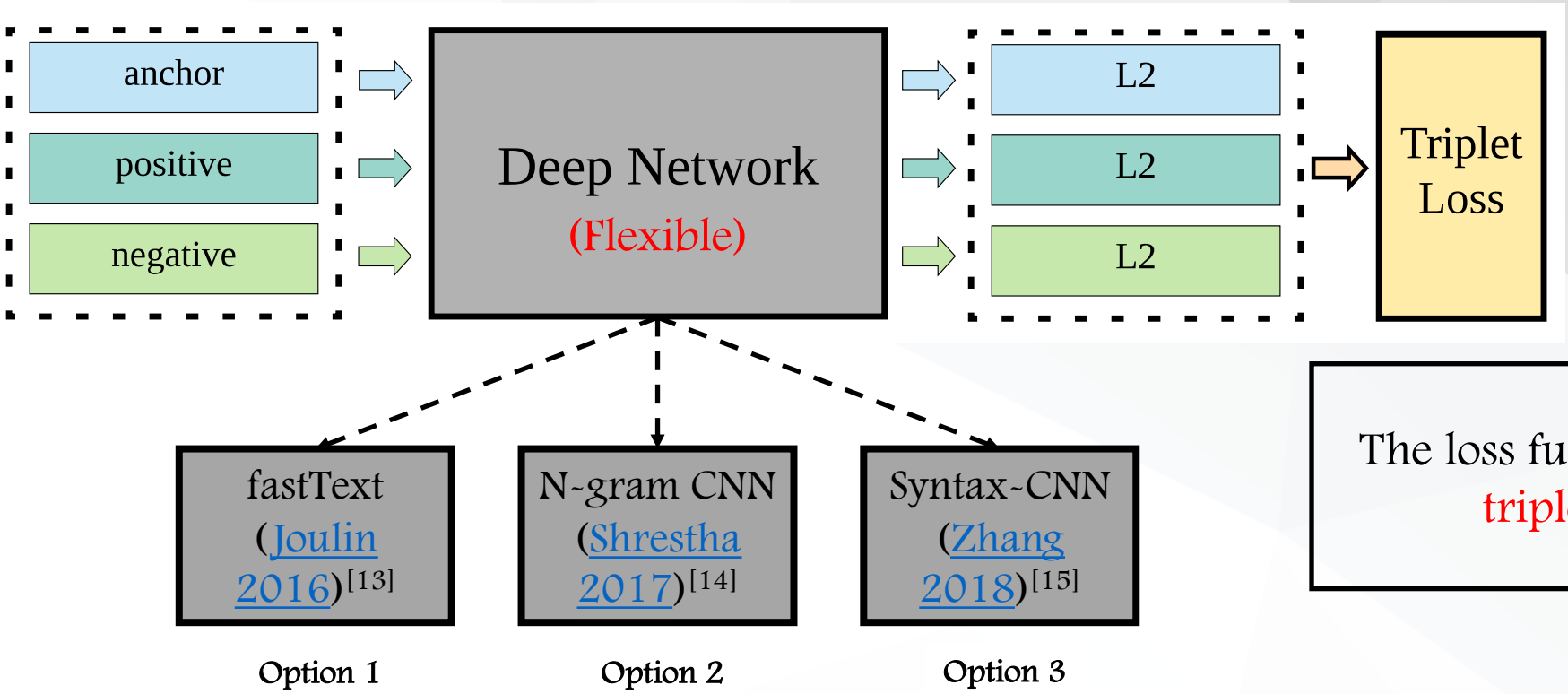
- $d_{ap} = ||\ f(x_i^a) - f(x_i^p)\ ||^2_2$: the squared Euclidean distance between an anchor and an positive.
- $d_{an} = ||\ f(x_i^a) - f(x_i^n)\ ||^2_2$: the squared Euclidean distance between an anchor and an positive
- α : a margin

$$L = \sum_{i=1}^{n} \max(0, \left|\left|f(x_i^a) - f(x_i^p)\right|\right|_2^2 + \alpha - \left|\left|f(x_i^a) - f(x_i^n)\right|\right|_2^2)$$

03 Method: Model Architecture

The input of the model is a triplet

The output of the model is three Euclidean embedding, which are processed by L2 Norm



The loss function is the triplet loss

03 Method: Triplet Selection

A triplet x_i is defined as $< x_i^a, x_i^p, x_i^n >$

Suppose there are 100 authors, and each author has 100 articles.

100 authors | 100 anchors | 99 positives | 99 * 100 = 9900 negatives

Total triplets: $n = 9{,}801{,}000{,}000$

It's IMPOSSIBLE to use all triplets to train the model!

03 Method: Triplet Selection

Random Selection

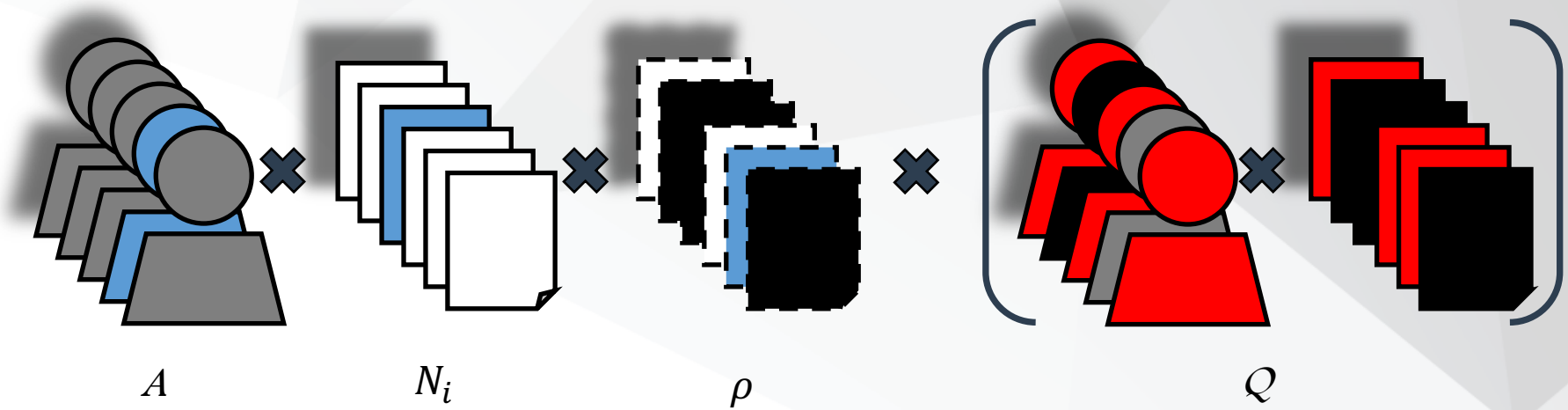
Simple, but not efficient for the model training.

A : total number of authors

N_i : total number of texts the author i has

ρ : a positive integer parameter (a cap on selected positives)

Q : a positive integer parameter (a cap on selected negatives)

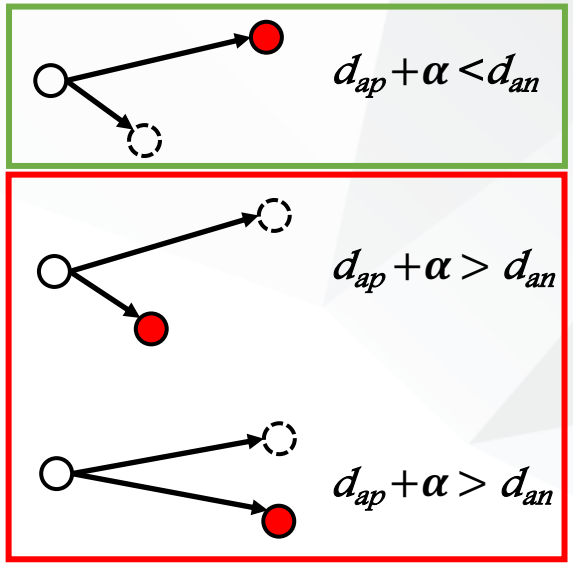


- For each author i, randomly select $P_i\langle anchor, positive\rangle$ pairs, where $P_i = N_i\rho$
- For each $\langle anchor, positive\rangle$ pair, randomly select Q negatives.
- Total number of triplets are capped under $\sum_{i=1}^{A} N_i\rho Q$.

03 Method: Triplet Selection

Dynamic Selection

Select efficient triplets before the beginning of each epoch



- An anchor x_i^a (the target text)
- A positive x_i^p (a text of the same author as the anchor)
- A negative x_i^n (a text of an author different from the anchor)

$d_{ap}=||f(x_i^a)-f(x_i^p)||_2^2$ $d_{an}=||f(x_i^a)-f(x_i^n)||_2^2$

Go to the next epoch

A effective triplet selection strategy can accelerate model training.

- Dataset is evenly and randomly divided into k partitions.
- For each partition, all $\langle anchor, positive\rangle$ are selected as RS
- For each $\langle anchor, positive\rangle$, randomly select Q negatives from candidates, each of which satisfies:

$$\left||f(x_i^a)-f(x_i^p)|\right|_2^2+\alpha\geq||f(x_i^a)-f(x_i^n)||_2^2$$

Baidu Security

Part-04
Experiments

04 Experiment: Dataset

- Crawled articles from 8 websites, such as huxiu.com, lieyun.com, tmtpost.com, etc.
- Removed incomplete and duplicate articles.
- Removed duplicate authors among multiple websites.
- Deleted author information in the article, like author name, WeChat ID, etc.

Website	Authors	Articles
renmin.com	609	2866
bjnews.com	591	4259
ynet.com	81	190
36kr.com	611	75412
cyzone.com	60	4136
tmtpost.com	748	20492
huxiu.com	827	9626
lieyun.com	340	16659
Total	3600	130000

04 Experiment

Again, Authorship Analysis contains three sub-problems

Authorship Attribution

The article's author is identified from a given list of candidates.

Authorship Verification

Whether the authors of two articles are the same or not.

Authorship Clustering

Articles of the same author are clustered.

04 Experiment: Preparation

Experimental Environment

- Tesla P40, 24GB
- Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
- RAM 42GB,
- Ubuntu 16.04.1
- Keras 2.0.9
- Tensorflow 1.4

Experimental Dataset

- Top N authors, which are sorted by the number of articles
- N is 5, 20, 50, 100, 500, 1000, 2000
- No more than 120 articles per author
- Train/Test = 8:2

	CN5	CN20	CN50	CN100	CN500	CN1000	CN2000
# authors	5	20	50	100	500	1000	2000
total # documents	600	2400	6000	12000	49447	63401	70132
max # docs per author	120	120	120	120	120	120	120

04 Experiment: Authorship Attribution(AA)

The article's author is identified from a given list of candidates.

Experimental Baselines:

CNN-NGRAM; CNN-WORD; NGRAM-SVM

Evaluation Metrics:

F1-micro-Score

Experiment Result:

Triplet-CNN > CNN-WORD & CNN-NGRAM, in all datasets.

NGRAM-SVM > Triplet-CNN, when the number of authors is 50 or 100.

Dataset	DS	RS	CNN-Word	CNN-NGRAM	NGRAM-SVM
CN5	**0.933**	0.933	0.9028	0.9208	0.785
CN20	**0.617**	0.581	0.551	0.559	0.557
CN50	0.443	0.402	0.373	0.393	**0.497**
CN100	0.419	0.32	0.282	0.292	**0.477**
CN500	**0.331**	0.236	0.154	0.177	0.276
CN1000	**0.296**	0.199	0.099	0.120	0.215
CN2000	**0.270**	0.188	0.079	0.085	0.189

(The best result on each dataset is in red)

Our method has obvious advantages than other baselines when the number of authors is large.

04 Experiment: Authorship Verification(AV)

Whether the authors of two articles are the same or not.

Experimental Baselines:

CNN-NGRAM; CNN-WORD; ALL-NGRAM; NOT-ALL-NEGRAM

Evaluation Metrics:

VAL (introduced by FaceNet)

Experiment Result:

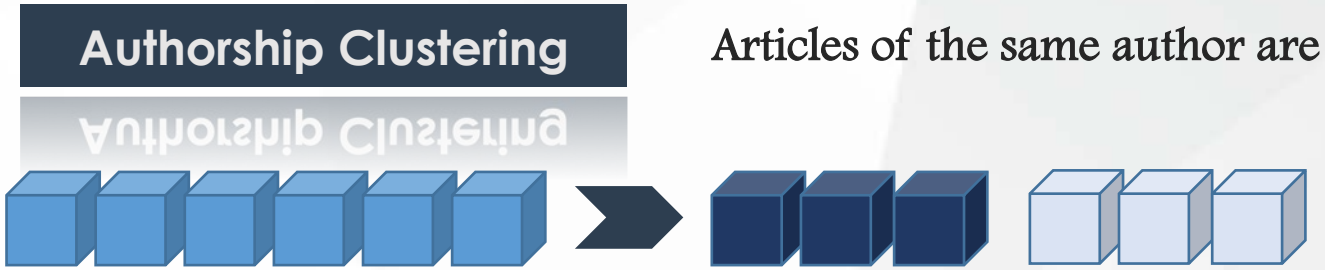
Triplet-CNN-DS > CNN-WORD, CNN-NGRAM, ALL-NGRAM, and NOT-ALL-NGRAM in all datasets.

Our method can calculate the similarity between articles more accurately

Dataset	DS	RS	CNN-Word	CNN-NGRAM	ALL-NGRAM	NOT-ALL-NGRAM
CN5	0.922	0.872	0.914	0.843	0.358	0.338
CN20	0.750	0.730	0.681	0.535	0.324	0.342
CN50	0.657	0.641	0.594	0.443	0.370	0.367
CN100	0.707	0.666	0.666	0.385	0.425	0.399
CN500	0.824	0.810	0.360	0.346	0.495	0.491
CN1000	0.832	0.827	0.332	0.329	0.495	0.493
CN2000	0.824	0.819	0.304	0.298	0.500	0.499

(The best result on each dataset is in red)

04 Experiment: Authorship Clustering(AC)



Articles of the same author are clustered.

Experimental Baselines:

LogEnt-HS (Hierarchical Clustering); LogEnt-KC (K-means Clustering)

Evaluation Metrics:

F-Bcubed score

Experiment Result:

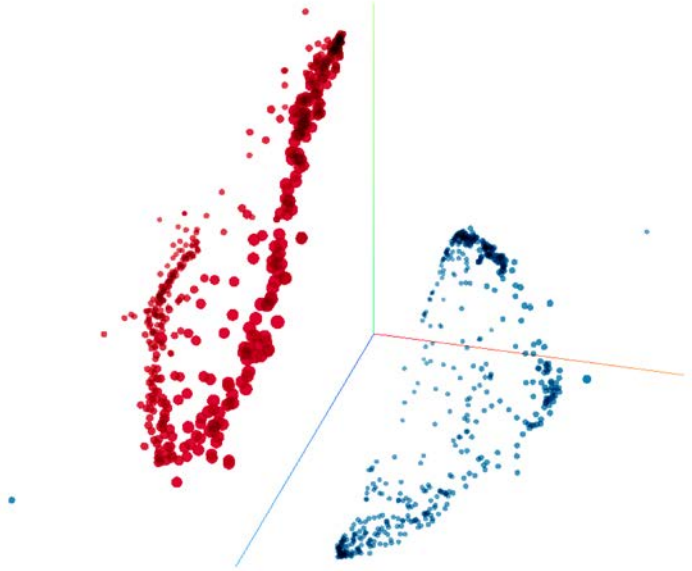
Triplet-CNN-DS > logEnt, in all datasets.

The advantage of Triplet-CNN-DS is more significant when clustering becomes more difficult.

Dataset	Triplet -CNN- DS-HC	logEnt- HC	Triplet- CNN- DS-KC	logEnt- KC
CN5	0.449	0.430	0.388	0.347
CN20	0.209	0.130	0.128	0.204
CN50	0.096	0.058	0.062	0.058
CN100	0.051	0.036	0.027	-

(The best result on each dataset is in red)

04 Experiment: Authorship Clustering(AC)



Triplet-CNN-DS

Visualization of clustering results

500 articles of the top two authors

Blue dots: Author1
Red dots: Author2

logEnt

Baidu Security

Part-05
Conclusion

05 Conclusion

Contributions

- We designed a triplet-based method, which is a unified embedding for authorship analysis, to track the fake news.
- We designed an efficient triplets selection strategy, which is proved be effective.
- We build the first Chinese dataset for authorship analysis.
- Experiments show that our method has better performance than other baselines, especially when the number of authors is large.

Future Works

- Testing our triplet-based method on more datasets.
- Optimizing the deep learning network and triplets selection strategy

Thanks!

liyiping02@baidu.com

Citation

- [1] Daily time spent on social networking by internet users worldwide from 2012 to 2019
- [2] False news stories are 70% more likely to be retweeted on Twitter than true ones
- [3] Inference & Disputed Authorship
- [4] A survey of modern authorship attribution methods
- [5] Computational Methods in Authorship Attribution
- [6] Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method
- [7] Authorship attribution in the wild
- [8] Authorship Attribution of Micro-Messages
- [9] Author Attribution with CNN's
- [10][14] Convolutional Neural Networks for Authorship Attribution of Short Texts
- [11][15] Syntax Encoding with Application in Authorship Attribution
- [12] FaceNet: A Unified Embedding for Face Recognition and Clustering
- [13] Bag of Tricks for Efficient Text Classification