General Fakeflow: Multi-domains Fake News Detection By Modeling The Flow Of Affective Information

Anthony TAING
anth248@stanford.edu, Department of Computer Science, Stanford University

CS224N-Winter 2022

OVERVIEW

- **Fake news detection** merely based on news content is tremendously challenging due to the usage of different text length, due to the variety of sources and different styles.
- **Binary Classification task**: predict if a news is true or fake.
- Previous works often use **a single source** like social media or **single topic** like Political, short texts.
- Previous works use **many features** (stylistic, ngrams…) or **external knowledge, social context information, don't use the emotions**.
- Fake news always play with **some affective factors** to manipulate the readers with **some eye-catching terms**. None have used the affective flow in texts.
- We aim to show the impact of **modeling the flow of this affective information** in a **cross-domains context with new datasets covering a wide range of topics.**

Data

3 Datasets with English content:

- **MultiSourceFake**: 5,994 real and 5,403 fake news, from online news websites.
- **ReCOVery**: 1364 true and 665 fake news about COVID-19 from 22 reliable and 38 unreliable websites and tweets.
- **Celebrity** : 250 true and 250 fake news about celebrities from web, magazines
- We split into 80% training and 20% test

References

[1]Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. FakeFlow: Fake news detection by modeling the flow of affective information. In the Association for Computational Linguistics(ACL), April 2021.
[2] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In the Association for Computational Linguistics(ACL), August 2018
[3] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for COVID-19 news credibility research. CoRR, abs/2006.05557, 2020.

APPROACH

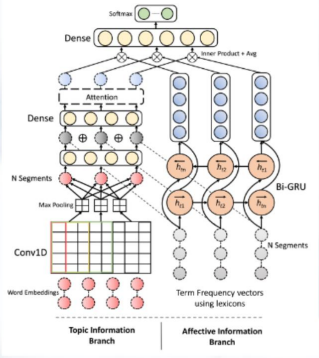


Figure 1: FAKEFLOW model

FAKEFLOW model:

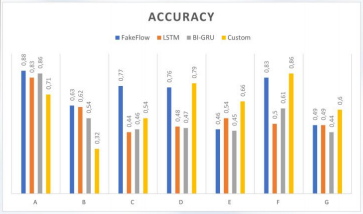
1) Topic Information branch
- Word2vec Embeddings
- Convolution+ Max pooling to learn features, highlight important words
- Fully connected layer
- We combine affective information and topic information into a fully connected layer to capture their interaction
- Self-Attention to capture the context of words

2) Affective Information branch
- **Term frequency features using Lexicons**: emotions changes(NRC lexicon), sentiment (positive/negative), morality (categories from Moral Foundations Dictionary), imageability (rated by their degree of abstractness), hyperbolic (high positive/negative).
- **Bidirectional GRU(Gated Recurrent Units)**
- **Final dot product and average +softmax**

Early stopping and ReduceLROnPlateau
Hyperparameters search: layers, dimensions, activations functions, learning rate, optimizer, pooling size, epochs, batch size.

Results and analysis

- FakeFlow still has better scores than baselines models
- FakeFlow has lower scores on unseen data, this is what we expected at the beginning of this project
- Generally lower scores with combined training data
- Our custom model outperforms all models when we combine data and in a cross-domain configuration.
- But weak with single training data



Data sets			FAKEFLOW	
	Training	Testing	Accuracy	F1score
A	80% MultiSourceFake	20%MultiSourceFake	**0.88**	**0.89**
B	80% MultiSourceFake	20% ReCOVery	0.63	0.75
C	80% MultiSourceFake	20% Celebrity	0.77	0.80
D	80% MultiSourceFake +80% Celebrity	20%Celebrity	0.76	0.78
E	80% MultiSourceFake +80% Celebrity	20%ReCOVery	0.60	0.44
F	80% MultiSourceFake +80% ReCOVery	20%ReCOVery	0.83	**0.89**
G	80% MultiSourceFake +80% ReCOVery	20%Celebrity	0.49	0.63

Table 2: Table of performances with Fakeflow model.

Data sets			LSTM	
	Training	Testing	Accuracy	F1score
A	80% MultiSourceFake	20%MultiSourceFake	**0.83**	**0.81**
B	80% MultiSourceFake	20% ReCOVery	0.62	0.75
C	80% MultiSourceFake	20% Celebrity	0.44	0.58
D	80% MultiSourceFake +80% Celebrity	20%Celebrity	0.48	0.54
E	80% MultiSourceFake +80% Celebrity	20%ReCOVery	0.54	0.67
F	80% MultiSourceFake +80% ReCOVery	20%ReCOVery	0.60	0.70
G	80% MultiSourceFake +80% ReCOVery	20%Celebrity	0.49	0.58

Table 3: Table of performances with LSTM model.

Data sets			Bi-GRU	
	Training	Testing	Accuracy	F1score
A	80% MultiSourceFake	20%MultiSourceFake	**0.86**	**0.88**
B	80% MultiSourceFake	20% ReCOVery	0.46	0.45
C	80% MultiSourceFake	20% Celebrity	0.54	0.37
D	80% MultiSourceFake +80% Celebrity	20%Celebrity	0.47	0.50
E	80% MultiSourceFake +80% Celebrity	20%ReCOVery	0.45	0.41
F	80% MultiSourceFake +80% ReCOVery	20%ReCOVery	0.61	0.27
G	80% MultiSourceFake +80% ReCOVery	20%Celebrity	0.44	0.32

Table 4: Table of performances with BI-GRU model.

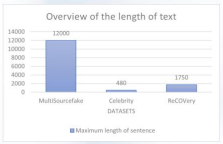
Data sets			Custom	
	Training	Testing	Accuracy	F1score
A	80% MultiSourceFake	20%MultiSourceFake	0.71	**0.83**
B	80% MultiSourceFake	20% ReCOVery	0.32	0.49
C	80% MultiSourceFake	20% Celebrity	0.54	0.70
D	80% MultiSourceFake +80% Celebrity	20%Celebrity	0.78	0.80
E	80% MultiSourceFake +80% Celebrity	20%ReCOVery	0.66	0.31
F	80% MultiSourceFake +80% ReCOVery	20%ReCOVery	**0.86**	0.80
G	80% MultiSourceFake +80% ReCOVery	20%Celebrity	0.60	0.61

Table 5: Table of performances with our custom model.

Weak with Celebrity dataset and unseen data:

- Default settings with a maximum sentence length of 800 words, but the maximum length of this dataset is 500 words, which is less than other datasets.
- 10 splits of texts into segments.
- Don't learn emotions with contracted words.

We have less cue words, emotions in each segment



METHODS

- **Metrics**: We used **Accuracy** and **F1 score** for the classification task.
- **Baseline**: We used **a BI-GRU model** and an end-to-end **LSTM baseline using Word2vec word vectors** like in the original FakeFlow method to predict labels.
- **Custom FakeFlow model**: replace all contracted terms (what's=what is…), replace Maxpooling with Maxpooling and Averagepooling, add dropout after each layer.

	Data sets	
	Training	Testing
A	80% MultiSourceFake	20%MultiSourceFake
B	80% MultiSourceFake	20% ReCOVery
C	80% MultiSourceFake	20% Celebrity
D	80% MultiSourceFake +80% Celebrity	20%Celebrity
E	80% MultiSourceFake +80% Celebrity	20%ReCOVery
F	80% MultiSourceFake +80% ReCOVery	20%ReCOVery
G	80% MultiSourceFake +80% ReCOVery	20%Celebrity

Table 1: Table of proportion of data in train and test sets.

Conclusion

- Fakeflow model is still efficient in cross-domains contexts, even if its performance decreased.
- But it is weak on shorter texts and unseen topics.
- We use only 10 segments(texts splits), and a maximum of 800 words.
- **Custom model** : more efficient when we combine training data in cross-domains
- **Future work: Try with other number of segments/maximum of words, Try other baselines, multiples languages**