Bert-Based Promotional Words Detection Classifier Development

Zikun Lin (supervided by: John R. Kender)

Columbia University, New York, NY, United States

Abstract

From the analysis last semester, we can see the necessity for us to find a way to get rid of advertising and promotion words from our corpora. After manually picking out promotion sentences, we get a list of ”blackwords” from video descriptions corpora, which can be used as a dataset in machine learning. In this report, I use the traditional model Word2Vec and the most advanced NLP model “BERT” combining with other machine learning methods to help us picking out the promotion sentences from video descriptions.

1. Introduction

As discussed in the previous report, there are a lot of promotion words and sentences appearing in the video descriptions. Typical categories of these words contains:

(a) Media promotion and subscription request. Such as: news, videos, subscribe, visit.

(b) URL links and social network account. Such as: twitter, facebook.

(c) Media name. Such as: vice, ‘arirang. In fact, the most “vice”s appear in the descriptions are not acting as “vice president”, but “VICE News” instead.

Before applying NLP and ML methods to video descriptions and transcripts dataset, we hope to get rid of these words as many as possible. However, applying manual work to all of these corpora is neither efficient nor doable. Therefore, we hope to develop a “blackwords-detection classifier” to assist us complete this task.

Several traditional NLP methods can take essential roles in our research. Especially in recent years, with the help of deep learning, models like LSTM, Word2Vec, BERT become popular in the NLP field. It enables us to convert words into vectors, making measurements and operations like clustering and classification possible. We can use these well-developed models and methods in our research to make the analysis more quickly and efficiently.

The main contributions of this research can be summarized as follows:

- I utilized the most recent context-based natural language processing model "BERT" to get embedding vectors for each word instance in our corpora.
- I utilized the "blackwords" dataset I collected last semester to train a "blackwords-detection classifier" based on SVM classifier. After grid search and fine-tuning, I achieved the accuracy of more than 60%. This classifier is able to be used in Stanley's deep cross-cultural system.
- I used video descriptions in AlphaGo and Florida shooting events as training set to train a final classifier. This classifier obtained a nearly 90% accuracy in the testing set, i.e. video descriptions in lunar rover and Thailand cave rescue events.

To help readers better understand the models used in experiments, I describe the datasets and some related work in Section 2. Then I will focus on my experiments in Section 4. And in Section 5, I will conclude my research and plan for the future.

2. Related Work

2.1. News Descriptions

As described in my previous report, in order to carry out the NLP part of our experiment, having clean data based on our task is necessary. In my experiments, I mainly use the news description data collected by Andy beforehand. These data contain paragraphs from the video description on several topics (AlphaGo, Florida shooting, lunar rover, Thailand cave rescue) in both YouTube and CGTN. For promotion sentences, I use the dataset manually picked by myself last semester, which contains two topics (AlphaGo and Florida shooting).

2.2. Word2Vec

Representing words in a vector space is an efficient way to group similar words and analyze the distribution of a set of words. Rumelhart et al. (1988), Mikolov et al. (2013a) and Mikolov et al. (2013b)'s papers described methods and improvements to represent word and phrases and their compositionality on a vector space. Particularly, Mikolov et al. (2013a) introduced the Skip-gram model, which is an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data, and it is one of the most popular ways to train word vectors.

In order to carry out our experiments quickly, I use Google's pre-trained word and phrase vectors[1], so that we do not need to take much time training from massive datasets. Instead, with the help of Řehůřek and Sojka (2010)'s Gensim library, we only need to call

```
model = gensim.models.KeyedVectors.load_word2vec_format()
```

function to load the model and get the vector representation we need.

2.3. BERT

In Nov.2018, Devlin et al. (2018) from Google AI Google AI open sourced a new technique for NLP pre-training called **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, or BERT.

BERT is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. Especially, with the "bidirectional" strategy, BERT is able to embed words according to its context. In order to utilize this strategy, they use the

[1] https://code.google.com/archive/p/word2vec/

straightforward technique of masking out some of the words in the input and then condition each word bidirectionally to predict the masked words.

With the help of their pretrained model, we are able to train models on our corpora in several hours just on our own machines. In my experiments, I used their most recently released "BERT-Large, Uncased (Whole Word Masking)" model[2], with 24 layers and 1024 hidden dimensions on their Github repo [3]. After minor modifications to their code (shown in Appendix 1), we can easily get a 1024-dimension embedding for every word occurrence in our training corpus.

2.4. SVM

Support **V**ector **M**achine (SVM) is a supervised learning model. The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. After about 30 years, Boser et al. (1992) introduced kernel SVM to create nonlinear margins. The current standard incarnation (soft margin) was proposed by Cortes and Vapnik (1995).

SVM can greatly help our task by providing a lightweight but effective classifier to classify two different types of words ("black" and "white"). In my experiments, I mainly use the SVM model implemented in sklearn Python package. In this model, we are able to adjust the kernel coefficient for radial-based-function (gamma) and penalty parameter of the error term (C).

[2] https://storage.googleapis.com/bert_models/2019_05_30/wwm_uncased_L-24_H-1024_A-16.zip

[3] https://github.com/google-research/bert

3. Hand-labeled blackwords

3.1. Video Descriptions on AlphaGo

3.1.1. Blacklist Words for Video Descriptions in English

Sentences containing advertisement and promotion are treated as “blacklist sentences”. The typical “blacklist sentence” is like this one[4]:

> *Subscribe to VICE News here: http://bit.ly/Subscribe-to-VICE-News*
> *Check out VICE News for more: http://vicenews.com*
> *Follow VICE News here:*
> *Facebook: https://www.facebook.com/vicenews*
> *Twitter: https://twitter.com/vicenews*
> *Tumblr: http://vicenews.tumblr.com/*
> *Instagram: http://instagram.com/vicenews*
> *More videos from the VICE network: https://www.fb.com/vicevideo*

The top 60 words in the blacklist after eliminating punctuation are shown in Table 1.

Word	Frequency	Word	Frequency
news	110	with	22
to	98	is	22
the	90	out	22
on	88	in	21
	77	a	21
subscribe	70	channel	20
vice	60	your	19
us	53	this	17
official	53	cnbc	17
for	52	website	16
and	50	arirang	16
here	46	by	16
httpwwwfacebookcomarirangtvtwitter	39	my	16
httptwittercomarirangworldinstagram	39	httpbitlysubscribetovicenewscheck	15
visit	36	httpvicenewscomfollow	15
of	35	herefacebook	15
‘arirang	34	httpswwwfacebookcomvicenewstwitter	15
news’	34	httpstwittercomvicenewstumblr	15
pagesfacebooknews	34	httpvicenewstumblrcominstagram	15
httpwwwfacebookcomnewsariranghomepage	34	httpinstagramcomvicenewsmore	15
httpwwwarirangcomfacebook	34	network	15
httpinstagramcomarirangworld	34	httpswwwfbcomvicevideo	15
more	31	like	14
facebook	29	tv	14
from	27	intel	14
youtube	26	please	13
you	26	at	13
videos	25	cbs	13
our	25	app	12
twitter	24	software	12

Table 1: Top 60 Words in English Blacklist on AlphaGo

From this blacklist, we can see several typical categories of words:

[4] https://www.youtube.com/watch?v=8dMFJpEGNLQ

(a) Media promotion and subscription request. Such as: news (110), videos (25),subscribe (70), visit (36), please (13).

(b) URL links and social network account. Such as: httpwwwfacebookcomarirangtvtwitter (39), facebook (29).

(c) Media name. Such as: vice (60), ‘arirang (34). In fact, the most “vice”s appear in the descriptions are not acting as “vice president”, but “VICE News” instead.

This list will help us a lot in future work dealing with other video descriptions and transcripts since there are many similar patterns in videos from other topics.

3.1.2. Blacklist and Whitelist Words for Video Descriptions in Chinese

Sentences containing advertisement and promotion are treated as “blacklist sentences”. The typical “blacklist sentence” in Chinese is like this one[5]:

> 螃蟹科技微信公众号：螃蟹科技（*pangxiekeji*）螃蟹科技 *QQ* 群 *419859745*
> 如果对我们的栏目有什么建议或者对智能数码有什么需要了解的，在公众号中回复你想了解的，我们来帮你解答。
> ***Translation:*** *“Crab Technologies” Wechat Official Account: Crab Technologies (pangxiekeji) “Crab Technologies” QQ Group 419859745. If you have any suggestions for our column or need to know about smart digital, reply what you want to know in the Wechat official account, we will answer.*

And the typical “whitelist sentence” in Chinese is like this one[6]:

> 柯洁将在下月迎战谷歌旗下的著名人工智能围棋软件 *AlphaGo*。
> ***Translation:*** *Ke Jie will be battling with Google’s famous AI Go software AlphaGo next month.*

As described in Section **??**, the cultural differences on descriptions make it more difficult for analysis in Chinese. In this particular topic on AlphaGo, as we can expect, besides the sentences most relevant to AlphaGo and Ke Jie (blacklist sentences) and the promotion and advertising sentences (whitelist sentences), there are also other sentences that don’t belong to any of these two lists, which are referred as “irrelevant sentences”. Although time-consuming, it is quite interesting to read through all of these Chinese descriptions. Some typical irrelevant sentences are shown below.

The following one[7] is collected from a “technology news weekly digest” video. AlphaGo only serves as a small part in this video. So most contents in this video are irrelevant with AlphaGo and Ke Jie.

[5] https://www.bilibili.com/video/av4116312
[6] https://www.iqiyi.com/v_19rrbbtzbw.html
[7] http://v.qq.com/page/u/m/2/u0305zm4lm2.html

三星 Note6/7 工程图曝光联想 Moto Z 真机图泄露柯洁 AlphaGo 即将开战
10 万块军工级手机发布全球首款带夜视仪的手机发布
Translation: *Exposure of Samsung Note6/7 Engineering Drawings, Lenovo Moto Z Real Machine Map Leakage, Ke Jie and AlphaGo are about to battle, 100,000 military grade mobile phones released, first mobile phone with night vision in the world*

The following one[8] is collected from a funny video imagining AlphaGo playing League of Legends game. These kinds of videos are not from the news, but there are several such kinds of videos on these Chinese websites.

153. 如果 AlphaGo 来玩英雄联盟
Translation: *If AlphaGo plays LOL*

The following one[9] is collected from an industry introduction sentence. It used "AlphaGo" to express that they are using the modern techniques and they are among the first tier.

英飞凌德累斯顿智能工厂，工业 4.0 的 "AlphaGo"
Translation: *Infineon Dresden Intelligent Factory, the "AlphaGo" of Industrial 4.0*

The following one[10] is a bit special. This is a self-edited video with no informative content, and there are several similar videos like this on Chinese video websites. The uploaders of these videos want to express their fondness for somebody or something, so they made these videos using the existing video footage. In this video, the content is mainly collected and edited from news video clips, so most scenes are relevant to the AlphaGo topic. Also, there are many keywords on this topic in the description. Therefore, it will be easily recognized as "related news" if using blacklists and whitelists only.

这个视频的构思想了一年多（是的没写错）从去年小李人机的时候开始想，直到今年才在小十一的古力……啊不是，鼓励之下开始动手
一个 *AI* 爱上了人类，最终他们在一起了故事 \# 严肃
第一次做剧情向，剧情比较凌乱，希望能看懂
送给小十一！希望喜欢！！注 *1*：*AlphaGo* 来自于 *Ex Machina-Domhnall Gleeson*
注 *2*：主 *CP* 为 *AlphaGo*/柯洁，副 *CP* 为木谷实/吴清源，古力/李世石
注 *3*：*2017* 年 *6* 月 *2* 日更新微调版本。具体剧情见回复
Translation: *This video has been conceived for more than a year (yes, correctly written) since Lee Sedol's battle last year, and it was not until this year that Gu Li was in eleventh ranking.*
This is a story. An AI falls in love with a human being and eventually they get together \# seriously
This is the first time that I make a story video, and the plot is messy. I hope you can understand it.
It's a present for the Eleventh! Hope you like it! Note 1: AlphaGo comes from Ex Machina-Domhnall Gleeson
Note 2: The main couple is AlphaGo / Ke Jie, secondary couples are Minoru Kitani / Wu Qingyuan, Gu Li / Lee Sedol.
Note 3: Updated fine-tuned version on June 2, 2017. See the reply for the specific plot.

[8] https://www.iqiyi.com/w_19rub12smp.html
[9] https://v.qq.com/x/page/i0188drze8u.html
[10] https://www.bilibili.com/video/av10975529/

In conclusion, Chinese descriptions are much more complicated, so it is challenging to carry out a two-class classification for Chinese descriptions.
After eliminating these irrelevant sentences, the top 60 words in the blacklist are shown in Table 2, the top 60 words in the whitelist are shown in Table 3.

Word	Translation	Frequency	Word	Translation	Frequency
，		87	科技	science and technology	10
的	's	64	更多	more	9
：		61	！		9
。		25	我	I	9
碧蓝	Azur	23	玩家	player	9
在	at	21	加入	enter	9
航线	Lane	20	了	have done ...	9
中途岛	Midway	16	qq		8
群	group	15	《		8
游戏	game	14	》		8
如果	if	14	com		8
微信	Wechat	14	…		8
集	episode	13	您	you	8
主	main	13	可以	can	8
alphago		13	喂	Hello	8
公众	public	12	id		8
号	account	12	服务器	server	8
交流	communicate	12	服	server	8
欢迎	welcome	12	644132397		8
都	all	11	请	please	7
allen		11	粉丝	fans	7
关注	follow	11	也	also	7
更	more	11	详细	detailed	7
up		11	攻略	strategy	7
有	have	10	尽	use all	7
围棋	Go	10	wiki		7
是	is	10	你	you	6
视频	video	10	对	to	6
、		10	和	and	6
大家	everyone	10	【		6

Table 2: Top 60 Words in Chinese Blacklist on AlphaGo

From the lists shown, we can see that the whitelist for Chinese is much more reliable than blacklist: For whitelist, there are about 10 words that appear more than 100 times, most of which are highly relevant to the topic. However, the top blacklist that has real meaning is “碧蓝 (Azur)”, which only has a frequency of 23. This phenomenon has shown that the blacklist in Chinese is much messy than whitelist, thus much less reliable.

[11] The word “dog” has the same pronunciation as the word “Go” in Chinese, so “Alpha Go” will sometimes be referred as “Alpha Dog” in Chinese news.

Word	Translation	Frequency	Word	Translation	Frequency
alphago		418	战胜	defeat	29
，		366	将	will do	29
的	’s	302	对弈	play chess with	29
柯洁	Ke Jie	177	狗	dog[11]	28
。		170	0		25
围棋	Go	131	棋手	chess player	25
大战	battle	117	deepmind		24
人机	human and computer	113	上	up	23
了	have done ...	106	4		23
“		89	比赛	game, competition	23
李世石	Lee Sedol	85	中	middle	23
”		84	谷歌	Google	22
人类	human beings	83	vs		22
在	at	82	第	-th	22
是	be	70	master		21
：		63	阿尔法	Alpha	21
人工智能	artificial intelligence	59	《		21
月	month	57	》		21
5		51	1		21
日	date	51	被	be done	20
3		44	用	use	20
与	and	43	不	no	20
战	battle	42	人	human	20
和	and	38	手	hand	19
中国	China	37	乌镇	Wuzhen (a place in China)	19
？		36	进行	be in progress	19
对	to	35	团队	team	19
、		31	我们	we	19
ai		30	你	you	18
！		30	马云	Jack Ma	18

Table 3: Top 60 Words in Chinese Whitelist on AlphaGo

3.2. Video Descriptions on Florida Shooting

After a discussion, we found that the AlphaGo event is a little general, which means several irrelevant events appeared under the search result of the “AlphaGo” keyword. In order to sort out better “blacklist” and “whitelist”, we turn our eyes to another event, the Florida Shooting tragedy.

3.2.1. Blacklist Words for Video Descriptions in English

The top 60 words in the blacklist after eliminating punctuation are shown in Table 4.

From the original data, we can observe that most of the videos come from different sources comparing to AlphaGo videos. However, they share many common blacklist words. For example, the English Blacklist on Florida Shooting shown in Table 4 and the English Blacklist on AlphaGo shown in Table 1 share 5 words in top 10, 8 words in top 20, 22 words in top 40. This means different media also use similar words in advertising and promotion. This observation makes our future work much easier since we can use the blacklists above to discover most of the targets in new topics.

Word	Frequency	Word	Frequency
news	1074	full	83
on	657	is	83
the	619	fox	78
and	511	local	77
cbs	501	episodes	74
to	400	google	74
here	362	all	74
of	324	cbc	72
nbc	234	our	69
subscribe	196	it	68
evening	182	as	67
you	164	broadcast	66
with	156	access	63
a	132	devices	61
morning	126	day	59
watch	121	stories	59
twitter	118	business	59
your	113	original	58
today	110	coverage	56
facebook	109	guardian	54
instagram	108	entertainment	53
	106	new	52
channel	105	video	52
for	105	digital	52
in	105	source	50
this	104	mobile	48
more	98	shows	47
live	97	breaking	46
latest	95	apps	45
from	91	across	45

Table 4: Top 60 Words in English Blacklist on Florida Shooting

4. Experiments

In this section, I will focus on my experiments leading to my final classifier and the results of these experiments.

4.1. Data Preprocessing

My previous report and Section 2.1 described several properties of promotional sentences in our corpora. Before using BERT model to get word embedding vectors, Several preprocessing steps are necessary. With the help of BERT vocabulary, I modified and added some of the steps from last semester's approach, which improves the quality of our corpora.

(1) Cleanup: For every word occurs in sentence, I used regular expression matching to separate words by spaces and several particular punctuations (!?,.:'"();). In order to get rid of the disturbances of multiple dots in urls, I also added another rule to detect and cleanup these urls.

(2) Tokenize: Not every word in the corpora is in the dictionary. Therefore, in order to carry out the word embedding successfully, we need to tokenize these words. There are several cases that we need to consider about:

- The word is in the dictionary: This is the simpliest case. We can directly use the original word as the token.
- The word is a combination of two other words in the dictionary: First several models of BERT carefully considered this case. They used a greddy algorithm to split up these kind of words. This will make our model much more complicated without performance improvement. Therefore, I did not apply this strategy in my experiments.
- The word is not in the dictionary: we use "[UNK]" token for every word not in the dictionary.

(3) (Optional) Remove stopwords and re-tokenize: Stopwords are those words that can be ignored in search engine. These words are necessary components in sentences, but they don't contribute to the overall topics or styles. We can remove or replace these words with a special token to ignore the effect of these words. In my following experiments, I replace these stopwords with a special token '#'.

As shown in Table 5, after preprocessing, we are able to get the number of words and unique words for each topic. For AlphaGo and Florida shooting topics, we are also able to get the number of promotional words and unique promotional words.

Topics	# Word Instances	#Promotional Word Instances
AlphaGo	23934	5534
Florida shooting	19826	6486
Lunar Rover	10280	?
Thailand Cave Rescue	21406	?

Table 5: Word Count

4.2. Word Embedding

In the previous report, I used word2vec model to embed every word to a 300-dimension vector space. Also, in order to plot the results on a two-dimensional figure, I also use the t-SNE method described in Maaten and Hinton (2008) to visualizes high-dimensional data by giving each data point a location in a two-dimensional map.

As described in 2.2 and 2.3, we can take advantage of two different word embedding methods developed by Google. Here are several main differences between them:

(1) Word2vec is a simpler model, and we can directly download the "word to vector" dictionary provided by Google and have a quick reference to get the vector.
BERT is a much more complicated model, and we need to calculate the vectors again for every word in new corpora.

(2) Word2vec is simply based on word. Regardless of its context, every instance of the same word share exactly the same word vector.
BERT is based on word and its context. The deep neural network in its model will calculate the word vector based on a vocabulary list as well as the "masked" context. This is beneficial for words with more than one meaning (e.g. bank).

Besides these reasons, BERT is more suitable for this projects in the following ways:

(1) Since our corpora only contain 200 video descriptions for each topic, it's obvious that we have very limited training examples. If using word2vec, the size of the training set will be limited to around 2,000, since the same word is treated as only one single sample. However, if using BERT, each word instance is a sample, which enlarges the size of the training set to more than 10,000. This can avoid over-fitting problem if we apply machine learning methods to it.

(2) We are training a binary classifier for each word, and there are many words that appear in both "blackword" list and "whiteword" list if we take every different word as a single sample. However, if we use BERT, every word instance is a single sample,

In recent years, BERT has been proved to be one of the best model in NLP field. It has already been applied to many NLP tasks like reading comprehension, cloze, etc. Although fine-tuning the model requires huge amount of computing resources and time, the original model itself is sufficient for our project. In the following sections, I will be using their latest model “BERT-Large, Uncased (Whole Word Masking)” published in May, 2019, which produces an 1024-dimension vector for each word instance.

Here is the comparison of word embeddings using word2vec (shown in Figure 1) and BERT (shown in Figure 2). Red dots represent “blackwords” and blue dots represent “non-blackwords”.

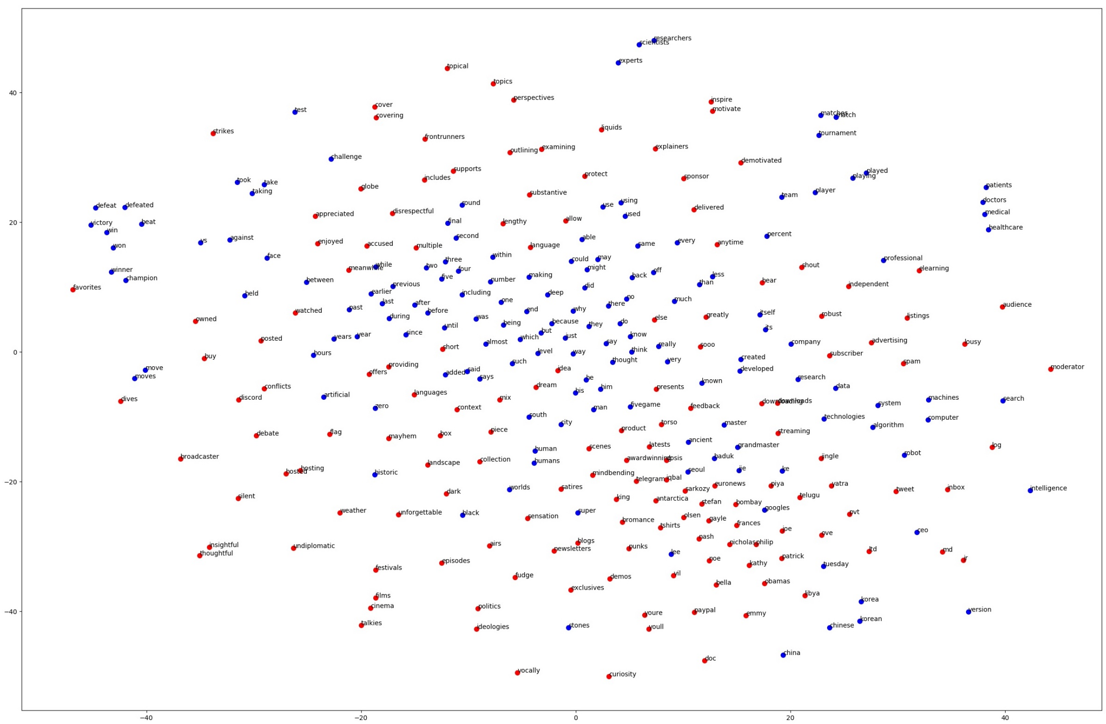

Figure 1: Word2vec embedding

From the figures shown, we can make several observations:

(1) Using word embeddings generated by word2vec, it’s difficult to observe a clear boundary based on the t-SNE figure. Since the overall boundary of Figure 1 is not so clear, I do these steps to help generate the “boundaries” shown in the Figure 3: I first use t-SNE as before to plot the words on a 2-d vector space. After that, I use SVM (support vector machine) to do the “classification” step. This step does not mean to “train a classifier” for future words. It just serves as a method to find the boundary between the two classes.

(2) Using word embeddings generated by BERT, the boundary becomes much clearer. We can also observe many clusters. In fact, every cluster usually represent the same word

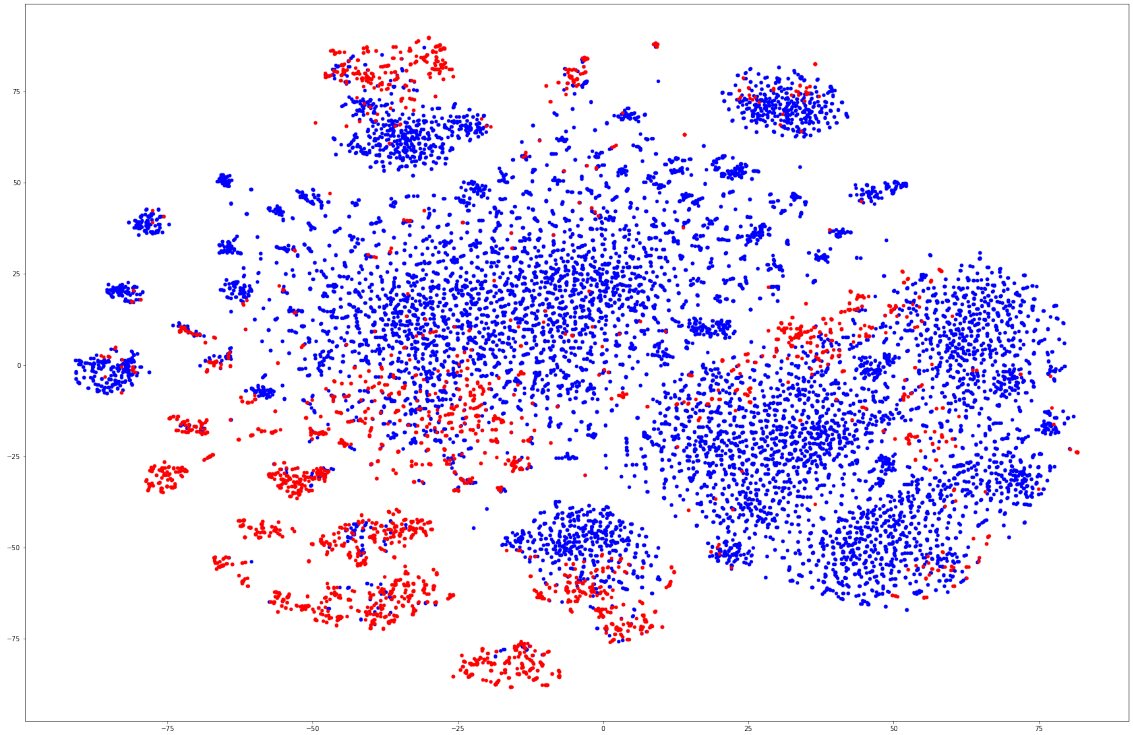

Figure 2: BERT embedding

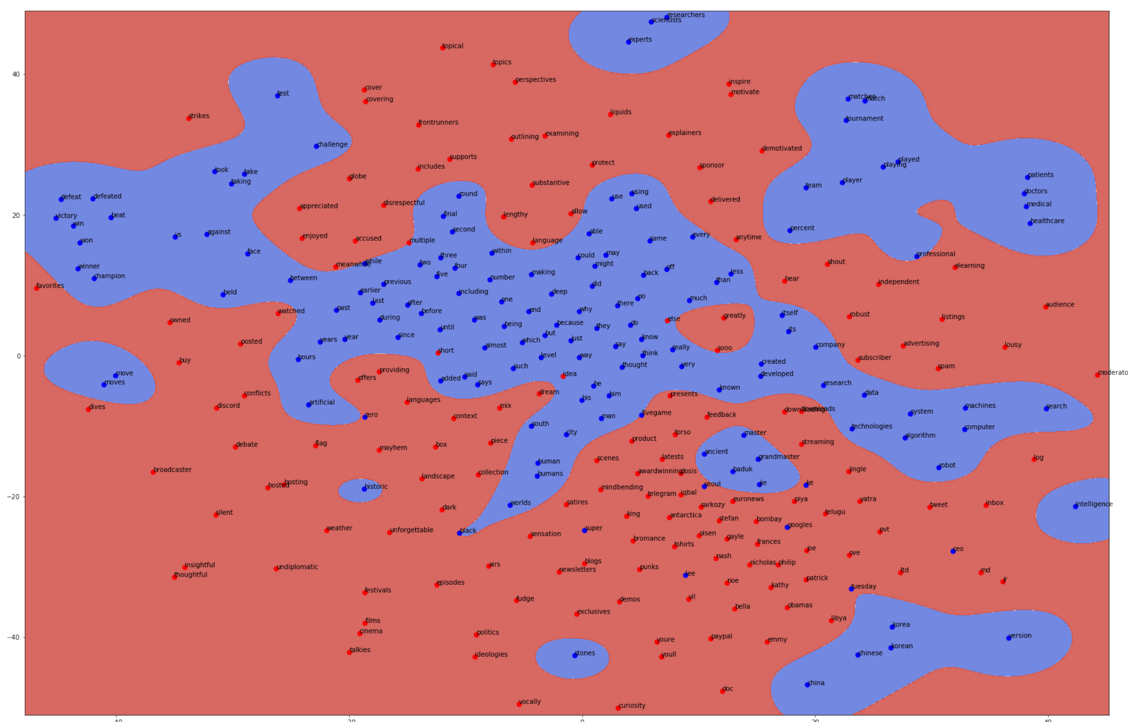

Figure 3: Word2vec embedding with boundaries generated by SVM

appearing in different places. If the two instances of the same word have similar usages or context, they tend to have similar word embedding.

4.3. Classifier Training

The BERT embedding shown in Figure 2 suggests that it's possible to train a classifier based on the word vectors. In this section, I will introduce the procedures of fine-tuning the model and the results I get from it. Since the word2vec embeddings are not able to provide us with good results, I will mainly focus on BERT embeddings in the following experiments and results.

4.3.1. Feasibility

In order to quantify the results and conclusions shown above, I used SVM with default C value and gamma=0.001 to train the very first model based on 70% of AlphaGo descriptions. The training error rate is 2.87% and 8.20%, which are both really good for this task. This result suggests that it's feasible to apply SVM model to train and fine-tune our classifier.

4.3.2. Grid Search and Fine-Tuning

In order to get the best performance, I used grid search to fine-tune the gamma and C values. In order to carry out the grid search, I trained 100 different models based on 10 gamma values and 10 C values, and then plot the results in a 3-dimensional figure.

The true positive rates based on different gamma-C pairs are shown in Figures 4 and 5.

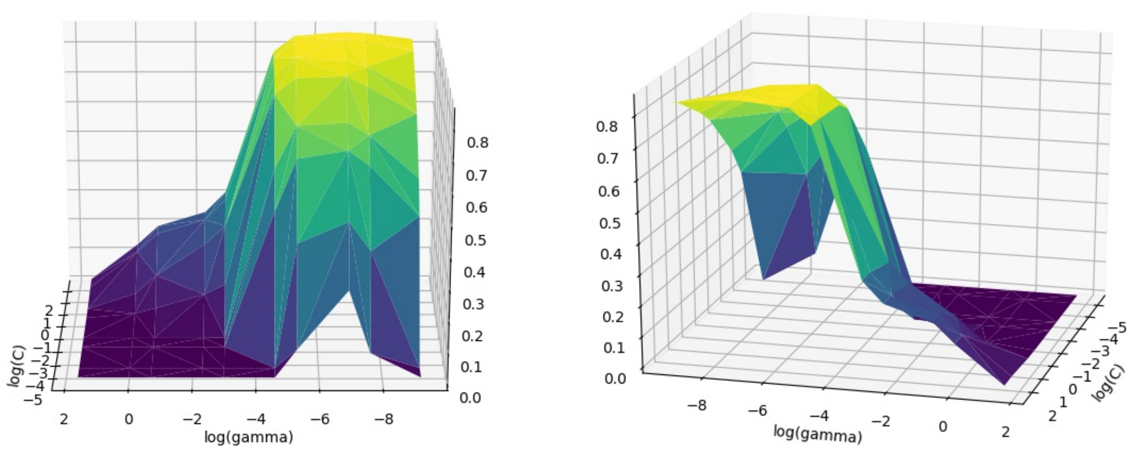

Figure 4: The true positive rates based on different gamma-C pairs on validation set (30% AlphaGo data)

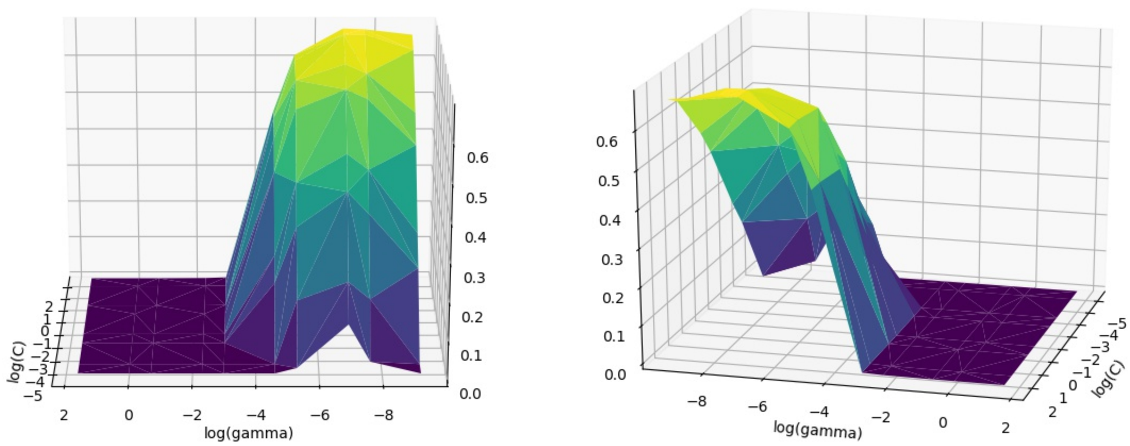

Figure 5: The true positive rates based on different gamma-C pairs on test set (Florida data)

From the two graphs, we can observe that with a lower gamma and higher C, we tend to get better performance. The basic reasons are[12]:

- Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning “far” and high values meaning “close”. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.
  Therefore, in our model, if using a lower gamma value, we are having higher σ in the original SVM model, leading to more support vectors. This is beneficial to our model with more than 15,000 training examples in only 1024 dimensions.

- The C parameter trades off correct classification of training examples against maximization of the decision function’ s margin. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words “C” behaves as a regularization parameter in the SVM.
  Therefore, in our model, since we have much more training examples than dimensions, we’d like to use a higher C to avoid “underfitting” problem.

Finally, I decided to use C=10 and gamma=0.0005 in my final model. This is based on the results above as well as reasonable analysis. In the following sections, I will use these parameters to train the final models.

[12] https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

4.3.3. Results on AlphaGo

For the AlphaGo dataset, I used 70% of sentences as training set, 30% of sentences as validation set, all Florida sentences as testing set. The results are shown in Table 6.

	Train	Validation	Test
Error Rate	1.10%	7.23%	25.10%
True Positive Rate	96.91%	84.61%	68.31%
True Positive Count	3542	1578	13544
False Positive Count	72	229	323
False Negative Count	113	287	6282
True Negative Count	13059	5040	6163

Table 6: Results based on 70% of AlphaGo description sentences as training set

From this table, we can also observe that the performance on testing set is much worse than training set and validation set. This is because we are training on AlphaGo dataset but testing on Florida dataset. In order to generalize our model to different topics, I trained on both AlphaGo and Florida dataset in the next section.

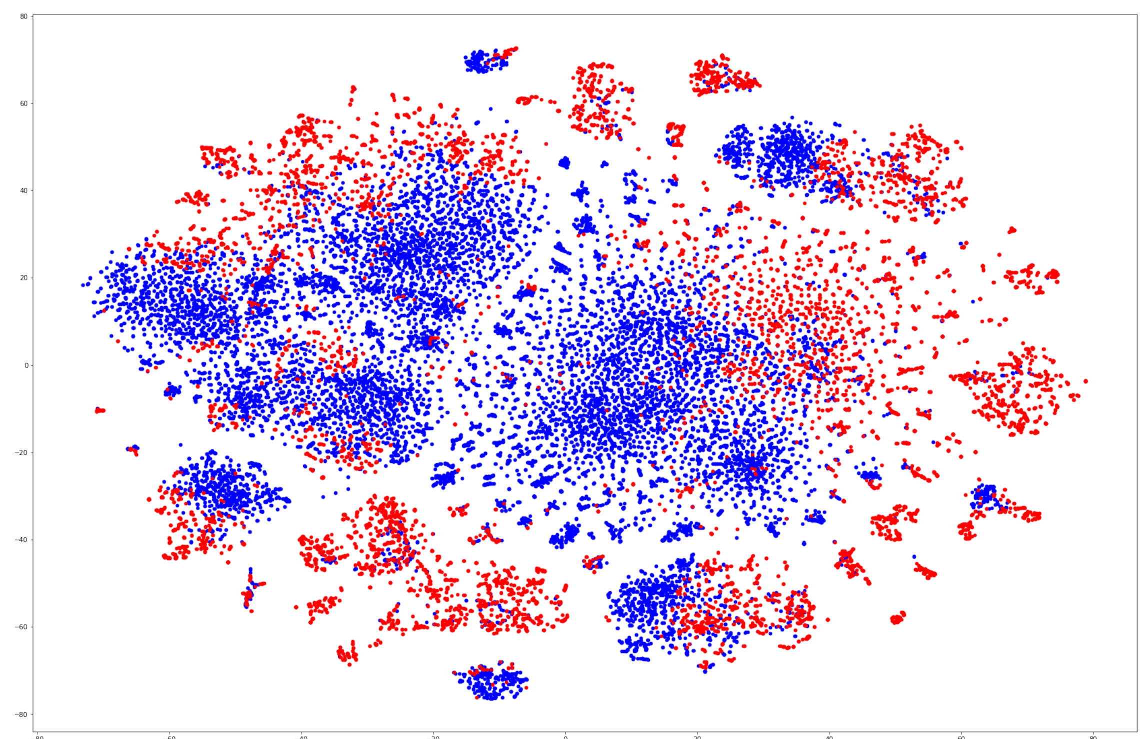

Figure 6: BERT embedding

4.3.4. Results on AlphaGo and Florida Shooting

Since we are mixing all AlphaGo and Florida description sentences together, we need to make sure that it's still possible to train an SVM model based on the dataset. Figure 6 shows the new BERT embedding after t-SNE operation.

Similar as previous experiments, I used 70% of sentences as training set, 30% of sentences as validation set. The prediction results are shown in Table 7.

	Train	Validation
Error Rate	1.96%	3.71%
True Positive Rate	97.96%	97.38%
True Positive Count	14037	10727
False Positive Count	405	252
False Negative Count	293	289
True Negative Count	20901	3328

Table 7: Results based on 70% of (AlphaGo+Florida) description sentences as training set

Comparing to the previous model, it's obvious that both accuracy and true positive rate improved a lot. And it shows that it's possible to use this model on other topics.

4.4. Final Results

After showing the possibility to train a model with high accuracy, I started use the model trained in the previous section to predict the other two topics, Thailand cave rescue and lunar rover.

After some observation, I found that "Thailand" topic contains a much higher ratio of blackwords than "lunar rover" topic. Therefore, I put the first 200 lines of the prediction in Appendix 2. We can observe that most of "blackwords" are predicted out, which means our model have a reasonable accuracy and sensitivity for promotional words detection.

5. Conclusion and Future Work

From the work described above, we can make several conclusions.

(a) We can manually collect "blacklist" and "whitelist" sentences and words from video descriptions. Due to the cultural differences, people treat descriptions differently on English and Chinese platforms. Also, video descriptions collected from Chinese video websites has more variety.

(b) As the most recent word embedding model, BERT has a much higher performance comparing to the traditional word2vec embedding. This allows us to train a model with more training samples based on the meaning in the contents.

(c) Using t-SNE, we are able to visualize the word vectors and observe the feasibility for training a classifier. Using fine-tuned SVM model, I finally trained a classifier with relatively high accuracy and true positive rate. This means we managed to use the hand-labeled data we have and BERT vectors to train a classifier for promotional words detection.

These are possible future steps for this research.

(a) Use BERT to train a new model for video descriptions in Chinese based on hand-labeled Chinese video descriptions on AlphaGo topic.

(b) Consider how to generalize this model to video transcript, which may have a different features and distributions comparing to video descriptions.

References

D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al., Learning representations by back-propagating errors, Cognitive modeling 5 (1988) 1.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013a).

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. http://is.muni.cz/publication/884893/en.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, ACM, pp. 144–152.

C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.

Appendix 1: Modification to the original BERT code

5.1. extract_features.py

```
@@ -31,6 +31,8 @@ flags = tf.flags

 FLAGS = flags.FLAGS

+flags.DEFINE_string("token_dir", None, "")
+
 flags.DEFINE_string("input_file", None, "")

 flags.DEFINE_string("output_file", None, "")
@@ -60,6 +62,10 @@ flags.DEFINE_bool(
     "Whether to lower case the input text. Should be True for uncased "
     "models and False for cased models.")

+flags.DEFINE_bool(
+ "do_whole_word_mask", False,
+ "Whether to use whole word masking rather than per-WordPiece masking.")
+
 flags.DEFINE_integer("batch_size", 32, "Batch size for predictions.")

 flags.DEFINE_bool("use_tpu", False, "Whether to use TPU or GPU/CPU.")
@@ -211,6 +217,8 @@ def convert_examples_to_features(examples, seq_length, tokenizer):
   """Loads a data file into a list of `InputBatch`s."""

   features = []
+
+ f = open(FLAGS.token_dir, 'w')
   for (ex_index, example) in enumerate(examples):
     tokens_a = tokenizer.tokenize(example.text_a)

@@ -279,15 +287,11 @@ def convert_examples_to_features(examples, seq_length, tokenizer):
     assert len(input_mask) == seq_length
     assert len(input_type_ids) == seq_length

- if ex_index < 5:
- tf.logging.info("*** Example ***")
+ if ex_index < 1000 and ex_index % 2 == 0:
       tf.logging.info("unique_id: %s" % (example.unique_id))
- tf.logging.info("tokens: %s" % " ".join(
- [tokenization.printable_text(x) for x in tokens]))
- tf.logging.info("input_ids: %s" % " ".join([str(x) for x in input_ids]))
- tf.logging.info("input_mask: %s" % " ".join([str(x) for x in input_mask]))
```

```
- tf.logging.info(
- "input_type_ids: %s" % " ".join([str(x) for x in input_type_ids]))
+ f.write(" ".join([tokenization.printable_text(x) for x in tokens]) + "\n")
+ f.write("\n")
+ tf.logging.info("***********")

    features.append(
        InputFeatures(
```

5.2. tokenization.py

```
@@ -172,7 +172,7 @@ class FullTokenizer(object):
    for token in self.basic_tokenizer.tokenize(text):
      for sub_token in self.wordpiece_tokenizer.tokenize(token):
        split_tokens.append(sub_token)
-
+ print("\t split tokens: " + ', '.join(split_tokens))
    return split_tokens

  def convert_tokens_to_ids(self, tokens):
@@ -327,35 +327,14 @@ class WordpieceTokenizer(object):

    output_tokens = []
    for token in whitespace_tokenize(text):
- chars = list(token)
- if len(chars) > self.max_input_chars_per_word:
- output_tokens.append(self.unk_token)
- continue

- is_bad = False
- start = 0
- sub_tokens = []
- while start < len(chars):
- end = len(chars)
- cur_substr = None
- while start < end:
- substr = "".join(chars[start:end])
- if start > 0:
- substr = "##" + substr
- if substr in self.vocab:
- cur_substr = substr
- break
- end -= 1
- if cur_substr is None:
```

```
- is_bad = True
- break
- sub_tokens.append(cur_substr)
- start = end
-
- if is_bad:
- output_tokens.append(self.unk_token)
+ #
+ # Zikun's new code
+ #
+ if token in self.vocab:
+ output_tokens.extend([token])
        else:
- output_tokens.extend(sub_tokens)
+ output_tokens.append(self.unk_token)
     return output_tokens


@@ -378,7 +357,7 @@ def _is_control(char):
   if char == "\t" or char == "\n" or char == "\r":
     return False
   cat = unicodedata.category(char)
- if cat.startswith("C"):
+ if cat in ("Cc", "Cf"):
     return True
   return False

@@ -390,8 +369,10 @@ def _is_punctuation(char):
   # Characters such as "^", "$", and "`" are not in the Unicode
   # Punctuation class but we treat them as punctuation anyways, for
   # consistency.
- if ((cp >= 33 and cp <= 47) or (cp >= 58 and cp <= 64) or
- (cp >= 91 and cp <= 96) or (cp >= 123 and cp <= 126)):
+ if cp == 91 or cp == 93:
+ return False
+ if (cp >= 33 and cp <= 47) or (cp >= 58 and cp <= 64) or (cp == 92) or \
+ (cp >= 94 and cp <= 96) or (cp >= 123 and cp <= 126):
     return True
   cat = unicodedata.category(char)
   if cat.startswith("P"):
```

Appendix 2: First 100 Result Segments on Thailand Rescue Topic (Bold words are predicted "blackwords")

hi friends , today would like share rescue operation plan cave incident (thailand) .
idea slightly different tesla ceo ' idea . think much possible rescue 12 boys coach .
idea insert **pipe** system cave rescue children help rescue pods (like pods used chile mine rescue operation) .
please watch video . also **welcome new** ideas suggestions helping take part rescue operation .
thank !
please share . **http : google . http : youtube .**
leader thailand ' rescue mission save 12 boys soccer coach flooded cave says , " limited amount time . "
spoke former thai navy seal , , died inside cave complex lack oxygen .
ben tracy reports chiang rai .
" cbs morning " channel : http : " cbs morning " : http
: latest installment " note self , " " cbs morning , " : http : " cbs morning " : http
: " cbs morning " facebook : http : . " cbs morning " twitter : http : " cbs morning " :
http : latest news best original reporting cbs news delivered . : http : news go ! download cbs news mobile apps : http : new episodes shows
love across devices next day , stream local news live , watch full seasons cbs fan favorites anytime , anywhere cbs access . try free ! http : king , "
cbs morning " offers thoughtful , substantive source news information daily audience 3 million viewers . emmy broadcast presents mix daily news , coverage developing stories national global significance , interviews
leading figures politics , business entertainment . check local listings " cbs morning " broadcast times .
one boys rescued caves thailand said " shocked " found .
12 boys football coach making first public appearance following ordeal caves .
please **http** :
thailand erupted celebration 12 youth football players coach trapped flooded cave northern chiang rai province two weeks rescued , following astonishing mission world .
final four school boys coach , trapped darkness cave complex 18 days , today carried operation came end .
residents chiang rai , city closest caves , took streets celebrate ,
drivers car horns pedestrians dancing outside hospital wild boar fc players recovering .
original article **: http :** . co . **original video : http : . co . daily**
mail **facebook : http : mail : http : mail snap : https : .** daily **mail twitter : http : mail : http : co . mail :**
https : **google . free daily mail mobile app : http : co .**
last member rescue team leave thai cave , australian doctor richard harris , lost father .
time http : **get closer world entertainment celebrity news time gives access insight people make watch , read share** . **https : youtube . ? list . money helps learn spend invest money . find advice guidance count negotiate ,**
save everything . https : **youtube . ? list . find latest developments science technology access brings ideas people changing world . https : youtube** .
? **list . let time show everything need** know drones **, autonomous cars , smart devices latest** inventions **shaping industries way : youtube** . ? **list . stay**

date breaking news around world trusted reporting , insight : youtube . ? list . connect : http : : https : : https : facebook .

: **https :** google **. : https : . ? : http : : time . brings insight , access authority news . news**

publication nearly century experience , coverage shapes understand world . daily news , interviews , science , technology , politics , health , entertainment , business updates ,

well exclusive videos person year , time 100 created acclaimed writers , producers editors . father australian doctor helped rescue trapped thai soccer team dies **:** youtube **.**

channel ’ teresa tang looks challenges divers face rescuing remaining 9 people trapped cave complex . meet wild heroes helped bring safety : **https :** us **:** https **:**

. com https : facebook . https : . https : twitter . https :

team divers resumed daring rescue mission free group boys flooded cave chiang rai , thailand july 9 . timeline happened far .

12 thai boys coach found alive inside cave complex chiang rai province , nine days went missing . ’ bird ’ view cave system . read full coverage

search rescue **operation : https : . follow us : https : . com https : facebook . https : . https : twitter . https :**

spokesperson us army team helping thailand rescue 12 boys assistant football coach flooded cave talk challenges met operation .

new details emerge 12 boys soccer coach survived two weeks flooded cave thailand . cbs **2** ’ anna werner reports .

thai official heading cave rescue says next phase operation start hours . joy benedict reports .

former elite diver navy died bringing oxygen flooded cave network youth football team coach still trapped .

: **http : . top stories today ? click watch :**

https : youtube **. ? list . : watched news channel ! http : youtube . available 13 languages : https : youtube . english :**

website : http : . : https : facebook . : http : : http : : http :

cave rescue operation trapped thai soccer team hit wall appears easy solution getting 12 boys coach safely .

welcome national , **flagship nightly newscast cbc** national **watch** videos

: https : youtube . voice opinion connect us online : national updates facebook : https : facebook . national updates twitter : https : national cbc television ’ flagship

news program . airing six days week , **show delivers news , feature documentaries analysis canada ’ leading journalists .**

indian played key role saving children trapped thailand ’ cave . prasad , designing engineer brothers limited company district , played key role saving lives 12 children coach

cave thailand . 12 . **hindi channel latest updates movies related videos . tube : https : youtube . follow us twitter : https : us facebook : https :**

facebook . circle google plus : https : google . download app : https : google .

patrick decker says ’ ” still long way go ” ongoing effort rescue members youth soccer team coach cave thailand .

members wild boar soccer team described moment found , **news** conference chiang rai , thailand . **nbc news : http : watch nbc video : http : news leading**

source global news information . find clips nbc nightly news , meet press , original digital videos . channel news stories , technology , politics , health , entertainment

, science , business , exclusive nbc investigations . connect nbc news online ! visit . com : http : nbc news facebook : http : nbc news twitter : http

: nbc news : http : nbc news : http : nbc news : http : soccer boys speak dramatic **rescue flooded cave nbc news**

cave system takes experienced diver five hours boys back full strength suffering exhaustion starvation found

. original article : http : . **co** . original **video :** http

: . co . daily mail facebook : http : mail : http : mail snap : https : . daily mail twitter : http : mail : http

: co . mail **: https :** google **. free daily mail mobile app : http : co .**

thailand ' art bridge chiang rai , created giant painting commemorate rescue operation 12 boys coach stuck cave . report nadia .

australian federal police divers , dr harris , whole australian team thank . term hero gets used lot . examples personnel worked cave rescue thailand .

pleasure morning . **australia** looks forward welcoming home safely **later week .**

listen tale race time rescue 12 boys soccer coach trapped cave two weeks . smarter . faster . colorful get **story http :** . even

? ! usa **youtube channel : http : usa today facebook : https : facebook . usa today twitter : https :**

thai navigate difficult terrain underground find alternative ways extract 12 boys football coach trapped cave complex since jun 23 . latest updates rescue efforts : **https : . follow us**

: https : . com https : facebook . https : . https : twitter . https :

captain dan brown discusses thailand cave rescue

divers working free 12 boys coach trapped cave northern thailand must navigate dark , flooded tunnels six hours reach . takes another five hours return . details extraordinary operation underway non

emerged thursday , pushed ahead multiple plans free boys trapped underground almost two weeks . 11 . **hindi channel latest updates movies related videos . tube : https : youtube . follow**

us twitter : https : us facebook : https : facebook . circle google plus : https : google . download app : https : google .

authorities thailand say twelve boys coach trapped inside cave ready dangerous dive flooded narrow passage . ' ve trapped almost two weeks . comments follow death former thai navy seal part

rescue team . died lack oxygen . john joe regan reports **. : http : : http : : http : : http : : http**

: **website : http : world**

two british divers , john jason honoured part rescuing 12 boys football coach cave complex thailand ' chiang rai province . meet wild heroes helped bring safety :

https **:** full coverage **rescue : https : us : https : . com https : facebook . https : . https : twitter . https** :

pressure mounting thai authorities bring forward rescue plan 12 boys coach trapped deep inside flooded cave northern thailand , death former navy diver drop oxygen levels underground . read : https :

rt . http : live http : rt ! http : youtube . like us facebook http : facebook . us telegram https : us https : us twitter http

: us http : us http : google . **(russia today)** global **news** network **broadcasting moscow** washington **studios . rt first news channel break 1 billion youtube views** .

john , bristol , one number foreign expert divers drafted rescue 12 boys football coach trapped thai cave nine days . tells bbc points west moment fellow

divers first discovered children alive . mr , member south mid wales cave rescue team , said knew found due smell cave . please **http** :

rescue ! ' wireless equipment used rescue 12 boys trapped caves thailand . **find** : **https : . follow us :**

facebook : https :
facebook . **: https : : https : .**
dive teams thailand rescued four boys flooded jungle cave monday . **watch** full episode **' world news tonight ' : https : full episodes world news tonight : http : go .**
northern thailand raced time ominous monsoon season youth soccer team trapped partially flooded cave advance heavy rains forecast later week .
two weeks trapped cave , 12 members wild football team coach rescued . risky operation led thai navy seals , international team managed get boys complicated often narrow
exit route . , coordinating operation , said ' thailand ' mission impossible ' **guardian news youtube http : cave rescue :** boys get **?** https
: **. support guardian https** : **guardian** https **: . guardian youtube network : guardian www . youtube . jones talks http : football http : sport http : culture http**
: science tech **http :**
elephant calf tumbled well eastern thailand reunited **mother** rescued group villagers . **us youtube : https : app apple store (ios**) **: https**
: **apple . download app google play (android) : https : google . follow us : facebook : https : facebook . : https :**
. **? : https : : https : . : http : . : http :**
thirteen members thai football team stranded cave rescued ending operation save . june 23rd , boys coach went explore cave complex , heavy rain flooded tunnels , leaving trapped
. john joe regan reports international rescue mission . **: http : : http : : http : : http : : http : website : http :**
world
may secret passage cave thailand youth soccer team (bottom right cave) trapped 11 days , emerged today . boys aged 11 16 told (left top
right) heard dogs barking , children playing despite 800 metres underground . led officials think may another way ' chimney hole ' surface . chiang rai provincial governor
, overseeing rescue , said 30 teams searching . believes must one boys able breathe long . left : trapped coach , 25