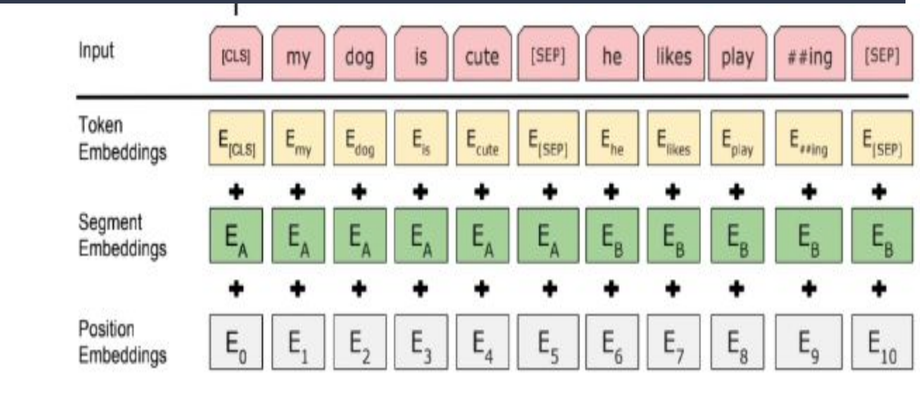
Evaluate BERT-wwm and XLM-R models for Chinese News Topic Prediction

Project Goal

体育	**黄蜂vs湖人首发：科比带伤战保罗 加索尔救赎之战 新浪体育讯北京时间4月27日，NBA季后赛首轮洛杉矶湖人主场迎战新奥尔良黄蜂，此前的比赛中，双方战成2-2平**
体育	1.7秒神之一击救马刺王朝于危难 这个新秀有点牛！新浪体育讯在刚刚结束的比赛中，回到主场的马刺通过加时以110-103惊险地战胜了灰熊，避免了让主场观众见证
体育	1人灭掘金！神般杜兰特！ 他想要分的时候没人能挡新浪体育讯在NBA的世界里，真的猛男，敢于直面惨淡的手感，敢于正视落后的局面，然后用一己之力，力挽狂澜
体育	韩国国奥20人名单：朴周永领衔 两世界杯国脚入选新浪体育讯据韩联社首尔9月17日电 韩国国奥队主教练洪明甫17日下午在位于首尔钟路区的足球会馆召开记者会，
体育	天才中锋崇拜王治郅 周琦：球员最终是靠实力说话2月14日从土耳其男篮邀请赛回到北京之后，周琦马上转机返回辽宁，由于之前比赛打得很辛苦，再加之时差的问题
体育	22+11!生涯最亮光辉 波什苦等8年终破首轮处男身新浪体育讯迈阿密热火主场97-91击败费城76人，总比分4-1淘汰对手晋级第二轮，即将迎战凯尔特人。波什登场40
体育	26+11+7热火杀神真被逼急了 若非他皇帝靠谁来拯救在勒布朗-詹姆斯状态稍显低迷的情况下，德维恩-韦德承担起来掌控热火进攻的重任。而韦德的表现也丝毫没有
体育	76人球员更衣室恶搞回应皇帝有些人连早餐都吃不完新浪体育讯北京时间4月28日(迈阿密时间4月27日)消息 与76人的第五战前，勒布朗-詹姆斯的一个比喻引起了很多
体育	ESPN为科比打抱不平 老鱼：只要有手有脚他就会上新浪体育讯北京时间4月28日，如果说季后赛首轮，湖人和黄蜂之间的比赛除了能看出卫冕冠军有点不给力之外。
体育	今日数据趣谈：阿杜比肩魔术师 热火中锋另类纪录新浪体育讯北京时间4月28日，NBA共进行了3场比赛。以下是今日比赛中诞生的一些有趣数据：凯文-杜兰特得到4
体育	从菩萨低眉到金刚怒目 石佛单节11分马刺找回自我新浪体育讯NBA季后赛首轮继续进行，在西区的焦点战役中，圣安东尼奥马刺主场历经加时赛以110-103击败孟菲
体育	八大1-3翻盘战役马刺找自信 奇迹从1.7秒压哨开始当还剩2.2秒时马努-吉诺比利的跳投被判为两分的时候，相信很多马刺球迷都已经开始绝望。犯规战术之后，马刺
体育	勇士宣布临时救火教练下课 斯马特升职一年便遭弃新浪体育讯北京时间4月28日，金州勇士今天通过官方宣布，不会选择执行主教练基斯-斯马特下赛季的球队选项，
体育	勇士弃帅不甘心却说软话 曝新帅候选阿联恩师在列新浪体育讯北京时间4月28日，据勇士官网报道，今天凌晨勇士火速宣布放弃了现任主帅基斯-斯马特，服务了金州
体育	半场3分皇帝屡遭争议判罚 满场裁判傻×为他鸣冤对于詹姆斯而言，他早已不在意去当篮球世界的王，MVP不是自己？不要紧，只要能和兄弟们一起拿下最后的总冠军
体育	广东半场狂飙狂胜山东 陈江华关键闪光稳住全队阵脚面对携四连胜而来的山东队，刚刚败退乌鲁木齐的广东东莞银行队昨晚凭借下半场的攻击狂潮，以129比92赢得
体育	即使出局大范仍执教魔术？ 高层：对主帅经理满意新浪体育讯北京时间4月28日，不管魔术首轮的最终命运如何，主帅斯坦-范甘迪和总经理奥蒂斯-史密斯看来都是高
体育	小牛vs开拓者前瞻VI：客场考验牛仔 拒绝拖入抢七新浪体育讯北京时间4月28日消息，波特兰开拓者队明天将回到主场接受达拉斯小牛队的挑战，双方的总比分目前
体育	希尔批活塞魔术治伤坑爹 麦蒂膝盖竟早有前车之鉴新浪体育讯北京时间4月28日消息，今天太阳老将格兰特-希尔参加FOX体育电台的节目，接受了主持人杰森-威尔洛
体育	惊!热火杀出三节20分小强 迈阿密的隆多今夜终现身新浪体育讯热火76人第五战的下半场，维持了整个季后赛的热火首发出现了一丝变化。常规赛末段因伤丢掉首发

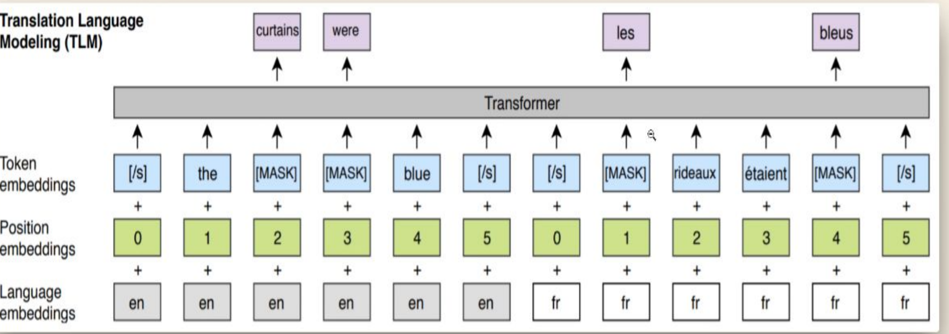
- 使用Bert-wwm和XLM-R对中文文本进行向量化
  - 使用Bert-wwm和XLM-R的原因
    - 都是跨语言模型
    - 都是使用的Wikipedia作为预训练的语料
- 使用Naive Bayes, Decision Tree, Random Forest等分类器对文本主题进行预测
- 观察每个主题预测的准确性，并分析原因
  - 数据包含：体育，娱乐，家居，房产，教育，时尚，时政，游戏，科技，财经 十种主题
- 比较分析两种模型在word embedding差异化的原因

Bert-wwm



- Whole Word Masking
  - Regular Bert: 原有基于WordPiece的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被mask。
  - Wwm: , 如果一个完整的词的部分WordPiece子词被mask，则同属该词的其他部分也会被mask。
- Example:
  - 使用语言模型来预测下一个词的probability。
  - 分词后：
    - 使用 语言 模型 来 预测 下 一个 词 的 probability 。
  - Regular Bert:
    - 使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 pro [MASK] ##lity 。
  - Bert-wwm:
    - 使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK] 。

XLM-RoBERTa



- 3种预训练任务
    - Casual language modeling(CLM)
    - Masked language modeling(MLM)
    - Translation Language Modeling(MLM)
- Translation Language Modeling:
    - 当无法使用英 语获取足够多预测mask的信息时，可以使用别的语言获取相关性
- 减少语种采样的bias，保证语料平衡

- 为什么使用XLM-R，而不是XLM
    - XLM中除了token embedding，position embedding还使用了language embedding
    - 而XLM-R中则放弃了language embedding这点和bert相同

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad with \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$$

Model script and Classification

```python
if __name__ == '__main__':
    parser = argparse.ArgumentParser()
    parser.add_argument("--type", type=str, default='train', required=True)
    args = parser.parse_args()
    _type = args.type

    assert _type in ['train', 'test', 'val']
    # _type = 'train'
    # labels = df['label'].unique().tolist()
    # labels_map = {label: index for index, label in enumerate(labels)}
    with open('./labels_map.json') as fp:
        # json.dump(labels_map, wp)
        labels_map = json.load(fp)

    xlmr = load_pre_trained_model()
    xlmr.eval()
    X = []
    Y = []
    data = read_data(_type)
    for row in tqdm(data):
        label, sentence = row.strip().split('\t')
        label_idx = labels_map[label]
        emb = sentence2vec(xlmr, sentence)
        if emb == None:
            continue
        emb = emb.tolist()
        X.append(emb)
        Y.append(label_idx)
    X = np.array(X)
    Y = np.array(Y)
    np.save(f'../embs/{_type}_X.npy', X)
    np.save(f'../embs/{_type}_Y.npy', Y)
```

```python
if __name__ == '__main__':
    """
    Main Function: change clf value for classification method
    """
    parser = argparse.ArgumentParser()
    parser.add_argument("--type", type=str, default='test', required=True)
    args = parser.parse_args()
    _type = args.type
    assert _type in ['test', 'val']

    X_train, Y_train = load_data('train')
    clf = LogisticRegression()
    clf.fit(X_train, Y_train)
    X_test, Y_test = load_data(_type)
    Y_test_pred = clf.predict(X_test)
    print(_type)
    print(classification_report(Y_test, Y_test_pred))
    print('\n')
    print(confusion_matrix(Y_test, Y_test_pred))
```

Classification Result

```
-----------------Random Forest--------------------------
-----------------Classification Report------------------
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      1000
           1       0.96      0.97      0.97      1000
           2       0.89      0.42      0.57      1000
           3       0.64      0.85      0.73      1000
           4       0.92      0.92      0.92      1000
           5       0.94      0.96      0.95      1000
           6       0.90      0.95      0.92      1000
           7       0.95      0.96      0.96      1000
           8       0.94      0.97      0.96      1000
           9       0.93      0.99      0.96      1000

    accuracy                           0.90     10000
   macro avg       0.91      0.90      0.89     10000
weighted avg       0.91      0.90      0.89     10000



-----------------Confustion Matrix----------------------
[[988   0   0   1   3   0   3   4   0   1]
 [  0 973   1   0   2   6   3  12   3   0]
 [  0  15 423 425  21  37  30  15   7  27]
 [  1   6  20 853  19  11  46   1   5  38]
 [  0   2  11  11 920   3  11   3  34   5]
 [  2  10  17   1   4 957   3   1   5   0]
 [  0   2   1  24  14   0 948   1   6   4]
 [  0   7   1   4  11   6   2 962   6   1]
 [  0   1   3   3   1   3   1  11 975   2]
 [  0   0   0   7   0   0   7   0   0 986]]
test
```

	Sports	Entert	Home	Real.E	Edu.	Fashion	Politics	Game	Tech.	Finance
Sports	992	0	0	1	1	1	2	2	1	0
Entert	0	979	3	2	3	4	0	2	5	2
Home	0	2	824	113	8	14	11	4	9	15
Real.E	2	2	27	878	8	10	34	0	2	37
Edu.	1	0	3	9	935	3	4	10	27	8
Fashion	0	4	4	0	4	986	0	0	1	1
Politics	0	1	2	22	9	0	955	1	6	4
Game	0	1	0	0	3	4	0	987	5	0
Tech.	0	0	1	1	1	5	0	5	983	4
Finance	0	0	0	7	0	0	4	0	0	989

Figure 1: Confusion Matrix for BERT-wwm

	Sports	Entert	Home	Real.E	Edu.	Fashion	Politics	Game	Tech.	Finance
Sports	982	4	0	1	2	2	4	3	0	2
Entert	0	961	2	0	5	12	2	11	5	2
Home	0	11	487	339	16	28	28	24	17	50
Real.E	0	6	22	864	19	8	33	0	4	44
Edu.	0	0	9	15	932	3	9	2	28	2
Fashion	1	6	8	0	4	973	2	2	3	1
Politics	0	2	0	19	17	0	941	1	8	12
Game	0	5	4	2	13	8	2	959	5	2
Tech.	0	1	3	3	0	9	1	13	969	1
Finance	0	0	0	14	1	0	3	0	0	982

Figure 2: Confusion Matrix for XLM-R

Classification Result

Method	Acc.	F1
BERT-wwm + LR	91.00%	0.9504
BERT-wwm + RF	90.00%	0.8976
XLM-R + LR	90.50%	0.9002
XLM-R + RF	89.77%	0.8914

Table 2: Experiment Results for Testset

Label	P	R	F1
Sports	0.9970	0.9920	0.9945
Entert	0.9899	0.9790	0.9844
Home	0.9537	0.8240	0.8841
Real.E	0.8500	0.8780	0.8637
Edu.	0.9619	0.9350	0.9483
Fashion	0.9601	0.9860	0.9729
Politics	0.9455	0.9550	0.9502
Game	0.9763	0.9870	0.9816
Tech.	0.9461	0.9830	0.9642
Finance	0.9330	0.9890	0.9602

Table 4: Prediction results of BERT-wwm with Logestic Regression

Label	P	R	F1
Sports	0.9990	0.9820	0.9904
Entert	0.9649	0.9610	0.9629
Home	0.9103	0.4870	0.6345
Real.E	0.6874	0.8640	0.7656
Edu.	0.9237	0.9320	0.9278
Fashion	0.9329	0.9730	0.9525
Politics	0.9180	0.9410	0.9294
Game	0.9448	0.9590	0.9519
Tech.	0.9326	0.9690	0.9505
Finance	0.8944	0.9820	0.9361

Table 5: Prediction results of XLM-R with Logestic Regression

Techniques Used

- 混淆矩阵
  - 查看是否过拟合/欠拟合
  - 查看每种类预测的结果
    - 如果预测错误，预测在哪里了

- 多种分类器模型
  - Logistic Regression, SVM, Decision Tree, Random Forest
  - 其中logistic regression和Random Forest准确率最高
  - 但同时Random Forest需要很久的运行 时间,

Conclusion and Report

Evaluate BERT and XLM-R models for Chinese News Topic Prediction

Luna Liu
yuli8896@colorado.edu
Matt Niemiec
matthew.niemiec@colorado.edu
Qiuyang Wang
qiwa8995@colorado.edu
Xinyu Jiang
xiji6874@colorado.edu

Abstract

Multi-lingual text classification gives brings a variety of challenges, but can have many benefits. Among the top-performing models in this field, BERT gives state-of-the-art results, and XLM-R, while less tested, shows great promise. We seek to analyze the two models side by side by training on and classifying Chinese news articles. Our findings reveal that BERT yields an impressive 95% accuracy, while XLM-R struggles to reach 91%. In addition to highlighting the importance of proper use of a model, one compelling reason for these results is BERT's robustness, even on medium-resource languages such as Chinese.

1 Introduction

News topic prediction is crucial for news industries around the world and society as a whole. For example, when the public health sector needs to communicate to the public, it is useful to be able to classify a news article so that the general public can readily access that information. Besides the application in public health, transforming unstructured data into distinct categories is also useful when compiling a list of relevant data and information on a subject. In the field of natural language processing (NLP), news topic prediction is always an important task. For example, a large volume of studies use news content to predict the trend in the stock market (Nikfarjam et al., 2010; Vargas et al., 2017), as well as monitor public health (Ng et al., 2020; Mahabaleshwarkar et al., 2019).

Furthermore, much of the data that exists in the world today is not in English, but another language. Therefore, it is important to be able to have NLP models that can analyze data in different languages, though this can prove challenging in more complex languages like Chinese. There is an abundance of data written in Chinese, and to make any use of it, it is important to analyze the top-performing models in Chinese text classification.

In this study, we explore text classification by implementing two pre-trained models, mBERT and XLM-R, on Chinese news articles. We first fine-tune each model on the training data. Then, we conduct a comparison by analyzing the performance of each model on the validation data. With these results, we compare the accuracy, F1 score, and confusion matrices of each model for deeper analysis.

2 Related Work

2.1 BERT and Chinese BERT-wwm

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is one of the highest-performing pre-trained models in NLP. The key technical innovation behind BERT is the bidirectional pre-training for language representations, which has helped overcome some of the previous limitations with standard, unidirectional pre-training. In order to conduct bidirectional pre-training, BERT uses two unsupervised tasks, masked language model (MLM) and next sentence prediction (NSP). These innovations greatly facilitate the fine-tune process, and has helped BERT achieve state-of-the-art performance for a large number of NLP tasks.

Recently, an updated version of BERT, BERT Whole Word Masking (BERT-wwm) was released. In the original BERT model, 15% of the WordPiece token was masked randomly in each sentence (Devlin et al., 2019). By contrast, BERT-wwm masks all WordPieces that belong to a whole word together (Cui et al., 2019). This is especially important for Chinese because most Chinese words are made up of several characters. Cui et al. (2019) then adapted the whole word masking strategy in Chinese BERT and re-trained the Chinese BERT

- Bert-wwn在文本分类上比XLM-R表现更加优秀
  - Bert-wwn 主要使用维基百科数据进行训练，故它们对正式文本建模较好
  - 在长文本建模任务上，例如阅读理解、文档分类，BERT和BERT-wwm的效果较好。
  - XLM的一个主要特点是使用多语言，信息互通进行预训练(BERT在多语种训练时信息不互通)，从而让模型能够掌握更多的跨语言信息
- 编写论文："Evaluate BERT and XLM-R models for Chinese News Topic Prediction"
  - Link:https://www.jianshu.com/p/f180fa4c0fe3