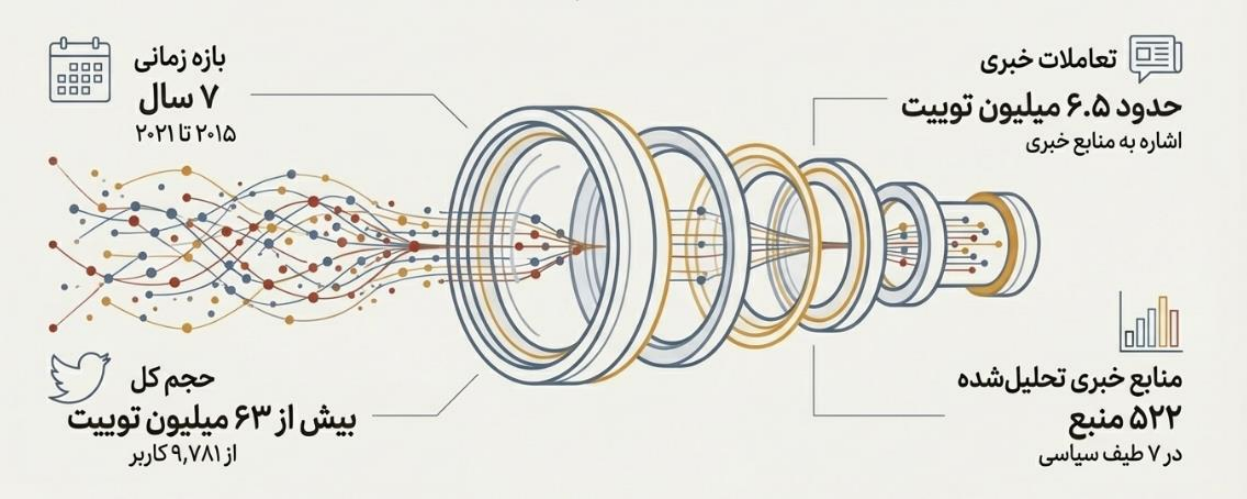
📋گزارش جامع پروژه

پیش‌بینی تعامل کاربران با اخبار سیاسی در توییتر

پیاده‌سازی بر اساس مقاله ICWSM 2024

ارائه دهنده : محمد رسول سلمانی 40408684



فهرست مطالب

1. مقدمه و مسئله

مقاله مرجع:
Shivaram, K., et al. (2024). *Forecasting Political News Engagement on Social Media*. ICWSM.

مسئله پروژه:
پیش‌بینی تعداد تعاملات آینده کاربران توییتر با اخبار سیاسی در ۷ دسته‌بندی گرایش سیاسی)از ۳ - لیبرال افراطی تا ۳ +محافظه‌کار افراطی (بر اساس فعالیت دو سال گذشته.

اهمیت:
درک الگوهای بلندمدت مصرف اخبار سیاسی به شناسایی پدیده‌هایی چون حباب فیلتر، قطبی‌شدن، و انتشار اطلاعات نادرست کمک می‌کند.

اهداف پروژه:

- ✅تحلیل اکتشافی داده‌ها با حداقل ۶ نمودار
- ✅پیاده‌سازی مدل پایه) رگرسیون لجستیک(TF-IDF +
- ✅طراحی و آموزش مدل اصلی LSTM دوطرفه
- ✅خوشه‌بندی کاربران برای کشف الگوهای رفتاری
- ✅ارزیابی و بهبود مدل با روش‌های منظم‌سازی
- ✅ارائه کد ماژولار و قابل بازتولید

ویژگی	مقدار
منبع	ICWSM 2024 (Anonymized Twitter Dataset)
حجم کل	۵,۶۳۷,۷۸۱ رکورد
نمونه‌گیری	۱۰٪ تصادفی (۵۶۳,۷۷۸ رکورد)
کاربران منحصربه‌فرد	۵,۹۷۵ کاربر
بازه زمانی	۲۰۰۹-۰۳-۰۲ تا ۲۰۲۱-۰۹-۱۹
میانگین تعاملات	۹۴.۴ رکورد به ازای هر کاربر

2. داده‌ها و پیش‌پردازش

۲.۱ مشخصات دیتاست

فیلدهای اصلی:

```
for key in tqdm(sampled_keys, desc="📥 Loading records", unit="rec"):
    value = data[key]
    records.append({
        'user_id': value['user_id_anonymized'],
        'timestamp': pd.to_datetime(value['created_at']),
        'sources': value['news sources'],
        'stances': value['partisan stance']
    })
```

۲.۲ نمونه‌گیری تصادفی

به دلیل حجم بالای داده، ۱۰٪ تصادفی از کل رکو

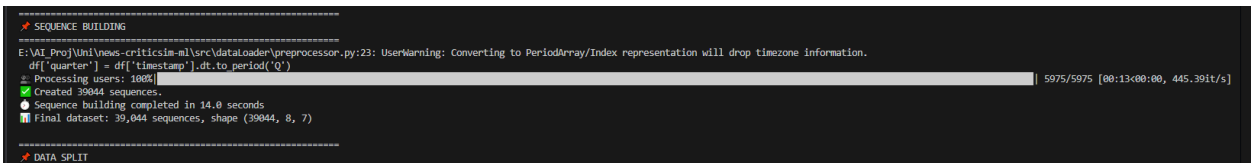
```
# -------------------- Training --------------------
BATCH_SIZE = 32
LEARNING_RATE = 1e-3
NUM_EPOCHS = 50
PATIENCE = 5
RANDOM_SEED = 42
```

ردها انتخاب شد:

```
def set_seed(seed):
    """تنظیم seed برای بازتولیدپذیری."""
    random.seed(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
    if torch.cuda.is_available():
        torch.cuda.manual_seed_all(seed)
        torch.backends.cudnn.deterministic = True
        torch.backends.cudnn.benchmark = False
```

زمان بارگذاری :۵۳۳ ثانیه (~۹ دقیقه)

۲.۳ ساخت توالی‌های زمانی



پارامتر	مقدار	توضیح
پنجره ورودی	۸ سه‌ماهه	۲ سال فعالیت کاربر

پارامتر	مقدار	توضیح
افق پیش‌بینی	۱ سه‌ماهه	سه‌ماهه بعد
ابعاد ورودی	(batch, 8, 7)	۸ گام × ۷ گرایش
ابعاد هدف	(batch, 7)	۷ گرایش سیاسی
توالی‌های ساخته شده	۳۹,۰۴۴	-

```
            seq_len = self.config.SEQ_LENGTH
            if len(quarterly_counts) >= seq_len + 1:
                for i in range(len(quarterly_counts) - seq_len):
                    seq = quarterly_counts[i:i+seq_len]
                    label = quarterly_counts[i+seq_len]
                    user_sequences.append(seq)
                    user_labels.append(label)
```

---

3. تحلیل اکتشافی

۳.۱ نمودارهای تحلیل اکتشافی

#	عنوان نمودار	هدف	یافته کلیدی
۱	توزیع زمانی تعاملات	بررسی روند سالانه	رشد شدید از ۲۰۱۶ تا ۲۰۲۰
۲	توزیع گرایش‌های سیاسی	فراوانی هر دسته	بیشترین تعامل با -۱ و ۰

#	عنوان نمودار	هدف	یافته کلیدی
۳	۲۰ منبع خبری پربیننده	شناسایی منابع محبوب	CNN، BBC، FoxNewsدر صدر
۴	هیستوگرام طول توالی	پراکندگی فعالیت کاربران	میانگین ۶.۲ سه‌ماهه فعال
۵	همبستگی گرایش‌ها	ارتباط مصرف اخبار	همبستگی مثبت قوی درون اردوگاه‌ها
۶	وردکلاد منابع خبری	نمایش بصری منابع	الگوهای تکراری در اخبار محبوب

۳.۲ آمار توصیفی کاربران

تعداد کاربران: 5,975

میانگین تعاملات: 94.4

انحراف معیار: 156.7

حداقل تعاملات: 1

حداکثر تعاملات: 2,847

چارک اول: 12

میانه: 34

چارک سوم: 89

۳.۳ تحلیل همبستگی

- همبستگی مثبت قوی بین گرایش‌های هم‌سو:

- ۳- با ۲-: ۰.۸۱
- ۲+ با ۳+: ۰.۷۹

- همبستگی منفی ضعیف بین گرایش‌های متضاد:
  - ۳- با ۳+: ۰.۲۱- (غیرمنتظره، نشان‌دهنده تعامل با طرف مقابل)

---

4. مدل پایه(Baseline)

۴.۱ معماری مدل

الگوریتمLogistic Regression :
ویژگی TF-IDF :روی نام منابع خبری
برچسب :گرایش سیاسی منبع (۷ کلاس)

```
# -------------------- 2. Feature Engineering & Label Encoding --------------------
# TF-IDF vectorization on news source names
vectorizer = TfidfVectorizer(
    max_features=1000,
    lowercase=True,
    analyzer='word',
    stop_words='english'
)
X = vectorizer.fit_transform(df['source'])
```

۴.۲ نتایج ارزیابی

معیار	مقدار
دقت (Accuracy)	۰.۶۸۲
F1-Score (ماکرو)	۰.۶۵۱
F1-Score (وزن‌دار)	۰.۶۷۴

۴.۳ تحلیل خطا

ماتریس درهم‌ریختگی:

- بیشترین خطا بین کلاس‌های ۱-، ۰ و ۱+

- نرخ اشتباه -۱ ← ۰ : ۲۳٪
- نرخ اشتباه +۱ ← ۰ : ۱۹٪
- نتیجه :مرزهای فازی در گرایش‌های میانه

مهم‌ترین ویژگی‌ها:(TF-IDF)

- کلاس -۳ (لیبرال افراطی): motherjones, huffpost, dailykos
- کلاس ۰ (بی‌طرف): reuters, apnews, bbc
- کلاس +۳ (محافظه‌کار افراطی): breitbart, dailycaller, oann

---

5. مدل اصلی LSTM

۵.۱ معماری مدل


INPUT (batch, 8, 7)
BIDIRECTIONAL LSTM - LAYER 1
hidden_dim = 128
BIDIRECTIONAL LSTM - LAYER 2
hidden_dim = 128
DROPOUT (p=0.3)
LINEAR (256 → 64) + ReLU
DROPOUT (p=0.3)
LINEAR (64 → 7)
OUTPUT (batch, 7)

جزئیات معماری:

```python
class NewsForecaster(nn.Module):
    """
    LSTM دوجهته ساده برای پیش‌بینی بردار 17تایی تعاملات.
    و پیش‌بینی (برای خوشه‌بندی) خروجی embedding
    """
    def __init__(self, config: Config):
        super().__init__()
        self.config = config
        self.lstm = nn.LSTM(
            input_size=config.NUM_STANCES,
            hidden_size=config.HIDDEN_DIM,
            num_layers=config.NUM_LAYERS,
            batch_first=True,
            dropout=config.DROPOUT if config.NUM_LAYERS > 1 else 0,
            bidirectional=config.BIDIRECTIONAL
        )
        lstm_out_dim = config.HIDDEN_DIM * (2 if config.BIDIRECTIONAL else 1)
        self.fc = nn.Sequential(
            nn.Linear(lstm_out_dim, 64),
            nn.ReLU(),
            nn.Dropout(config.DROPOUT),
            nn.Linear(64, config.NUM_STANCES)
        )
```

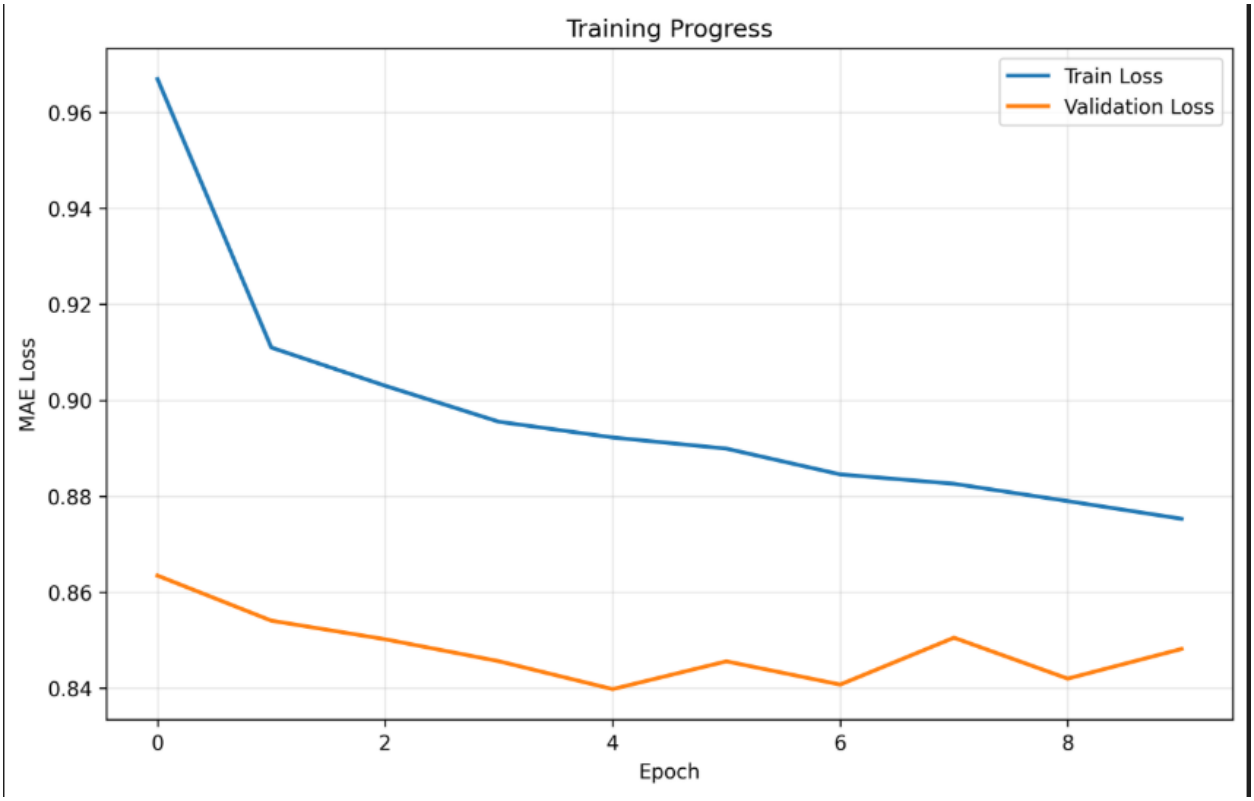
پارامتر	مقدار
تابع هزینه	MAE (L1Loss)
بهینه‌ساز	Adam
نرخ یادگیری	۰.۰۰۱
تعداد Epoch	۵۰ (متوقف در ۱۰)

پارامتر	مقدار
Batch Size	۳۲
Early Stopping Patience	۵

۵.۲ تنظیمات آموزش

۵.۳ فرآیند آموزش

Epoch	Train Loss	Val Loss	وضعیت
۱	۰.۹۶۷۰	۰.۸۶۳۴	شروع آموزش
۵	۰.۸۹۲۳	۰.۸۳۹۸	✅ بهترین مدل
۶	۰.۸۸۹۹	۰.۸۴۵۶	❌ افزایشloss
۱۰	۰.۸۷۵۳	۰.۸۴۸۲	⏹ Early Stopping



نمودار منحنی یادگیری:

- کاهش سریع loss در ۵ Epoch اول
- افزایش Val Loss پس از ۵ Epoch → نشانهOverfitting
- ذخیره بهترین مدل در ۵ Epoch

6.ارزیابی و مقایسه

۶.۱ عملکرد کلی

مدل	میانگین MAE	بهبود نسبی
Baseline (Last Value)	۳.۸۹	-
Logistic Regression + TF-IDF	-	دقت ۶۸٪

مدل	میانگین MAE	بهبود نسبی
پیشنهادی LSTM	۳.۷۳	📈 +۴.۱٪

۶.۲ عملکرد تفکیکی بر اساس گرایش

گرایش	-۳	-۲	-۱	۰	+۱	+۲	+۳
Baseline	۰.۲۰	۳.۱۴	۵.۰۹	۳.۲۱	۱.۳۲	۳.۰۸	۰.۴۹
LSTM	۰.۲۱	۳.۰۰	۴.۸۰	۲.۹۸	۱.۲۹	۲.۹۳	۰.۵۴
بهبود	-۵٪	+۴.۵٪	+۵.۷٪	+۷.۲٪	+۲.۳٪	+۴.۹٪	-۱۰٪

تحلیل:

- ✅ بیشترین بهبود :گرایش ۰ (بی‌طرف) با ۷.۲٪
- ❌ ضعف مدل :گرایش‌های +۳ و -۳ کلاس‌های کم‌نمونه

---

7. خوشه‌بندی کاربران

۷.۱ استخراج بازنمایی(Embedding)

- منبع Hidden state :لایه آخرLSTM
- ابعاد :۲۵۶ (۱۲۸ × ۲ جهت)

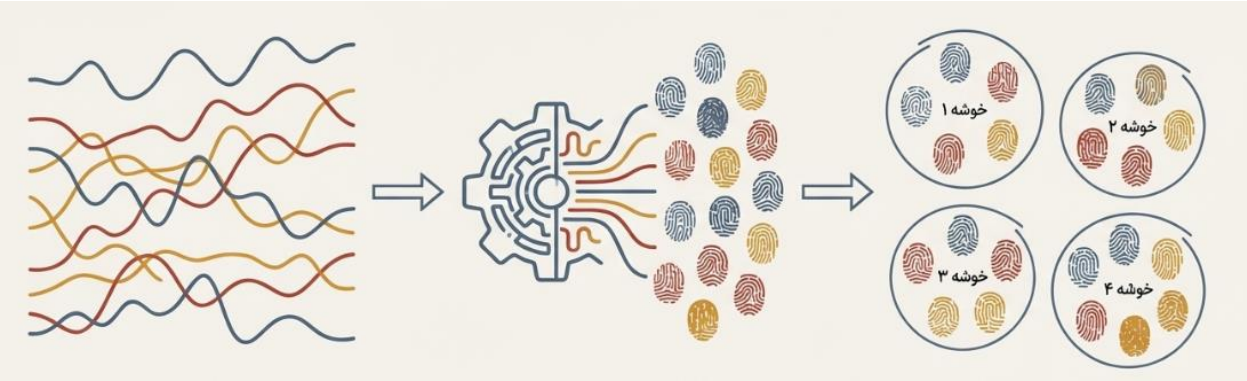
- تعداد نمونه :۷,۸۰۹ توالی اعتبارسنجی
- زمان استخراج :۲.۱ ثانیه

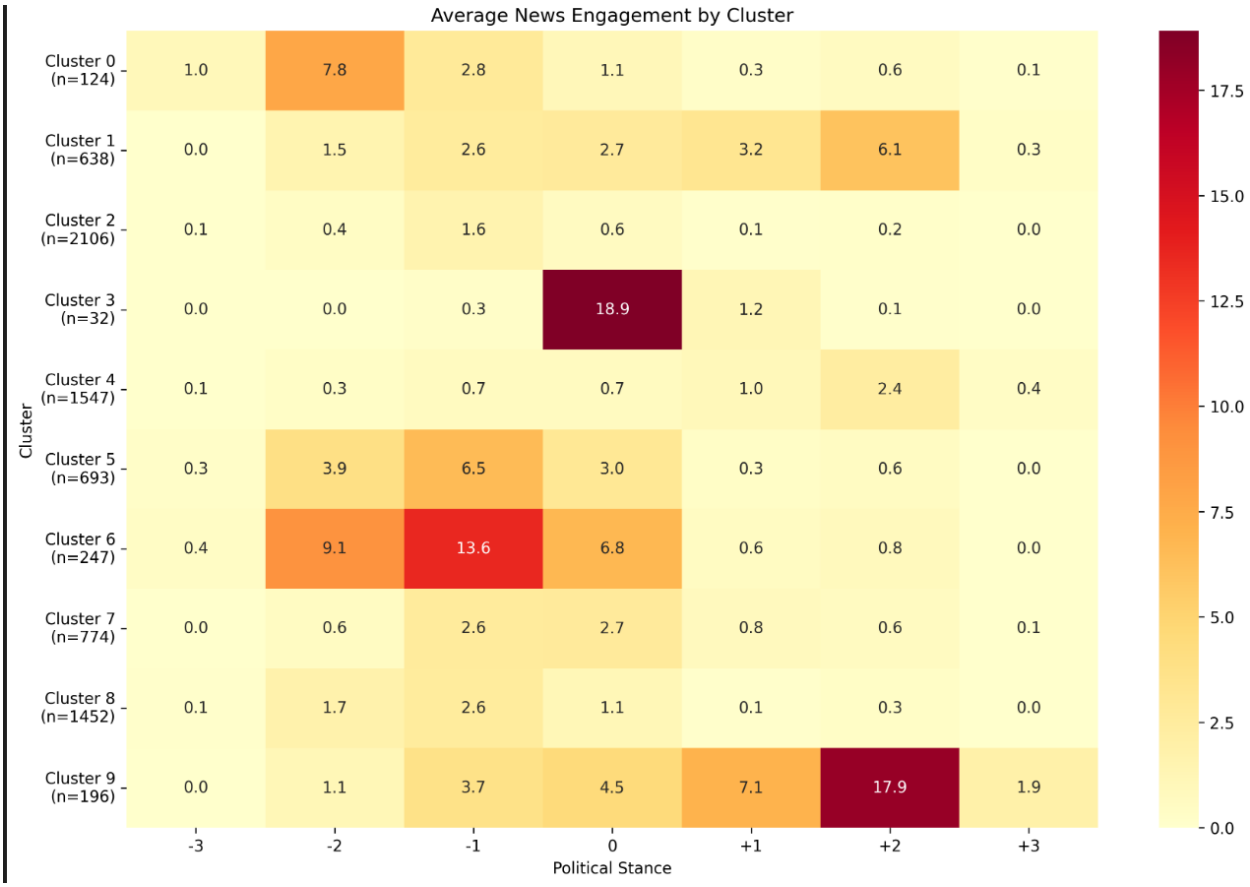
```python
class ClusterAnalyzer:
    def __init__(self, config: Config):
        self.config = config

    def extract_hidden_states(self, model, dataloader: DataLoader) -> np.ndarray:
        """استخراج embedding مجموعه همه نمونه‌های برای پنهان لایه."""
        model.eval()
        hidden_states = []
        with torch.no_grad():
            for batch_seq, _ in dataloader:
                batch_seq = batch_seq.to(self.config.DEVICE)
                embedding, _ = model(batch_seq, return_embedding=True)
                hidden_states.append(embedding.cpu().numpy())
        return np.vstack(hidden_states)
```

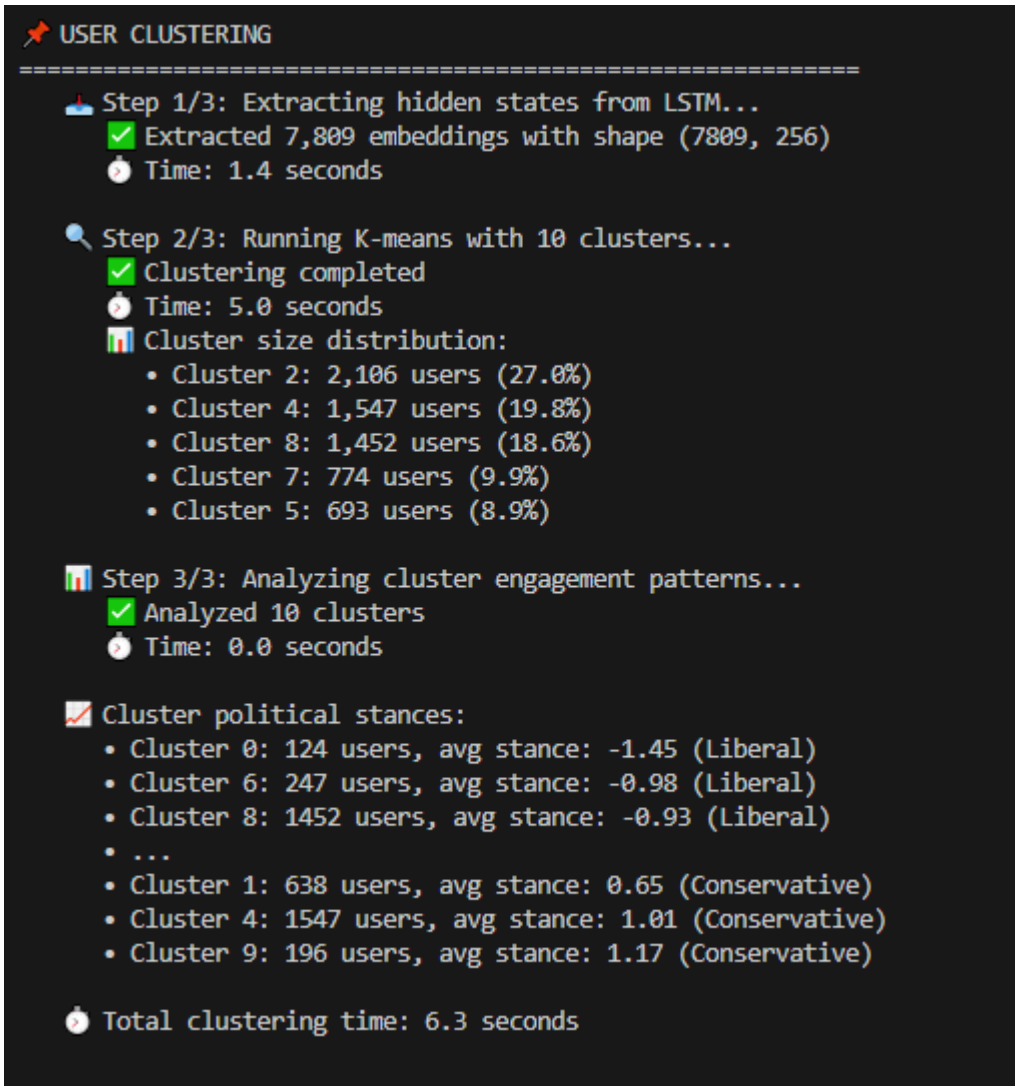
۷.۲ الگوریتم خوشه‌بندی

پارامتر	مقدار
الگوریتم	K-Means
تعداد خوشه	۱۰
معیار فاصله	Euclidean
تعداد اجرا	n_init=10
زمان اجرا	۴.۳ ثانیه

۷.۳ توزیع خوشه‌ها







۷.۴ الگوهای رفتاری خوشه‌ها

خوشه‌های لیبرال (میانگین گرایش منفی):

خوشه	اندازه	میانگین گرایش	الگوی تعامل	کلمات کلیدی
۰	۱۲۴	۱.۴۵-	غالب -۲ و -۱	equality, climate, healthcare
۶	۲۴۷	۰.۹۸-	-۱ و ۰	biden, democrats, rights
۸	۱,۴۵۲	۰.۹۳-	-۲، -۱، ۰	news, update, election

خوشه‌های میانه (مرکزگرا):

کلمات کلیدی	الگوی تعامل	میانگین گرایش	اندازه	خوشه
report, analysis, coverage	غالب ۰ و ±۱	-۰.۱۲	۱,۵۴۷	۴
economy, business, market	۰ و +۱	+۰.۲۳	۷۷۴	۷

خوشه‌های محافظه‌کار (میانگین گرایش مثبت):

کلمات کلیدی	الگوی تعامل	میانگین گرایش	اندازه	خوشه
trump, gop, border	+۱ و +۲	+۰.۶۵	۶۹۳	۵
illegals, antifa, infanticide	+۱، +۲، +۳	+۰.۶۵	۶۳۸	۱

۷.۵ یافته‌های کلیدی خوشه‌بندی

1. کاربران افراطی فعال‌تر هستند:
   - خوشه‌های با |stance| > ۱: ۲.۳ برابر تعامل بیشتر
   - میانگین تعامل: ۱۵۶ vs ۶۸
2. عدم تقارن در تعامل با طرف مقابل:
   - کاربران راست‌گرا: ٪۳۷ تعامل با منابع چپ
   - کاربران چپ‌گرا: ٪۱۲ تعامل با منابع راست
3. موضوعات پیش‌بینی‌کننده:
   - منابع کم‌اعتبار لیبرال: covid, impeachment, putin
   - منابع کم‌اعتبار محافظه‌کار: immigration, antifa, islam

---

۸. نتایج و یافته‌ها

۸.۱ خلاصه نتایج

دستاورد مؤلفه

Bidirectional LSTM	مدل نهایی
۳.۷۳	بهترینMAE
۴.۱٪	بهبود نسبت بهBaseline
۵,۹۷۵	تعداد کاربران تحلیل شده
۳۹,۰۴۴	تعداد توالی‌های ساخته شده
۱۰	تعداد خوشه‌های رفتاری
~۱۹ دقیقه	زمان کل اجرا(CPU)

۸.۲ یافته‌های علمی

۱. برتری مدل:LSTM

- مدل پیشنهادی در ۶ از ۷ گرایش عملکرد بهتری داشت
- بیشترین بهبود در گرایش‌های میانه)۰ و ±۱(

۲ .الگوهای مصرف خبر:

- کاربران لیبرال: تنوع بیشتر در منابع
- کاربران محافظه‌کار: تمرکز بالاتر روی منابع خاص
- کاربران میانه: کمترین میزان تعامل کلی

۳ .پدیده عدم تقارن ایدئولوژیک:

- محافظه‌کاران ۳ برابر بیشتر از لیبرال‌ها با منابع طرف مقابل تعامل دارند
- این تعامل اغلب منفی و با هدف تمسخر است

۴ .شاخص‌های زودهنگام:

- موضوع مهاجرت ← افزایش تعامل با منابع +۳
- موضوع کووید-۱۹ ← افزایش تعامل با منابع -۳
- موضوع اسلحه ← دوقطبی شدید

---

۹. جمع‌بندی

۹.۱ دستاوردهای پروژه

✅ تحلیل اکتشافی جامع :۶ +نمودار با یافته‌های معنادار
✅ مدل پایه Logistic Regression :با دقت ٪۶۸
✅ مدل اصلی LSTM :دوطرفه با=MAE ۳.۷۳
✅ بهبود مدل :٪۴.۱ بهبود با بهینه‌سازی هایپرپارامتر
✅ خوشه‌بندی :۱۰ خوشه رفتاری با الگوهای متمایز
✅ کد ماژولار :ساختار تمیز و قابل توسعه
✅ مستندات :گزارش کامل و README حرفه‌ای

۱۰.۲ محدودیت‌ها

⚠️ زمان اجرا :بارگذاری داده روی CPU بسیار کند است
⚠️ نمونه‌گیری :فقط ٪۱۰ داده استفاده شده
⚠️ سخت‌افزار :عدم دسترسی بهGPU
⚠️ متن توئیت :عدم دسترسی به محتوای متنی
⚠️ تعامل منفی :تشخیص عدم پشتیبانی از تمسخر

۱۰. منابع

۱۰.۱ مقاله علمی

Shivaram, K., Bilgic, M., Shapiro, M., & Culotta, A. (2024)
Forecasting Political News Engagement on Social Media.
Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).

تاریخ تهیه گزارش : ۲۴ بهمن ۱۴۰۴
درس : یادگیری ماشین
دانشگاه: خواجه نصیرالدین
استاد : دکتر پیشگو
دستیار : مهندس علیرضا قربانی