مشخصات مقاله

عنوان : Predicting Emerging Trends: A Machine Learning Approach to Topic Popularity on Social Media
نویسندگان : Zhe Wu et al.
مجله : PeerJ Computer Science (2025)

1.مسئله اصلی مقاله و اهمیت آن

رشد سریع شبکه‌های اجتماعی باعث تولید روزانه حجم عظیمی از محتوا و موضوعات (Topics) شده است. تنها بخش کوچکی از این موضوعات به «ترند» یا موضوعات محبوب تبدیل می‌شوند، در حالی که بسیاری دیگر به‌سرعت فراموش می‌شوند. مسئله اصلی مقاله این است:

آیا می‌توان در مراحل اولیه شکل‌گیری یک موضوع در شبکه‌های اجتماعی، پیش‌بینی کرد که آیا آن موضوع در آینده محبوب (Popular) خواهد شد یا خیر؟

اهمیت این مسئله از چند جنبه قابل توجه است:

- **کنترل و شناسایی زودهنگام اطلاعات نادرست و شایعات** که می‌توانند افکار عمومی را تحت تأثیر قرار دهند.
- **کاربردهای تجاری و بازاریابی** مانند زمان‌بندی کمپین‌ها و تشخیص موضوعات داغ بازار.
- **تحلیل افکار عمومی و سیاست‌گذاری** در بحران‌ها یا رویدادهای اجتماعی.

برخلاف بسیاری از کارهای قبلی که به مدل‌های سری زمانی متکی‌اند (و نیازمند داده‌های بلندمدت هستند)، این مقاله تلاش می‌کند **محبوبیت نهایی یک موضوع را تنها با استفاده از ویژگی‌های اولیه آن (مثلاً ۱۲ ساعت اول)** پیش‌بینی کند. این تغییر دیدگاه، مسئله را از «پیش‌بینی روند» به **«مسئله طبقه‌بندی زودهنگام»** تبدیل می‌کند که از نظر مهندسی و عملیاتی بسیار کاربردی‌تر است.

2.ورودی‌ها و خروجی‌های مدل / سیستم

ورودی‌ها

ورودی مدل، مجموعه‌ای از **ویژگی‌های استخراج‌شده در مراحل اولیه انتشار هر Topic** است. این ویژگی‌ها در چهار دسته کلی قرار می‌گیرند:

1. ویژگی‌های آماری پایه (طول متن، تعداد لینک، ایموجی و ...)

2. ویژگی‌های متنی و احساسی(Sentiment) ، نوع موضوع

3. ویژگی‌های زمانی (زمان انتشار، روز هفته، تعطیلات)

4. ویژگی‌های انتشار اولیه (تعامل کاربران در ۱۲ ساعت اول)

در مجموع، **۳۸ ویژگی** برای هر Topic تعریف شده است.

خروجی‌ها

خروجی سیستم یک **برچسب دودویی (Binary Classification)** است:

- Topic → 1محبوب(Popular)
- Topic → 0غیرمحبوب(Non-popular)

معیار محبوبیت بر اساس یک امتیاز نهایی محاسبه‌شده از تعداد پست‌ها، لایک‌ها، کامنت‌ها و اشتراک‌گذاری‌ها تعیین می‌شود. اگر این امتیاز از یک آستانه مشخص (مثلاً ۲۵۰۰) عبور کند، Topicمحبوب تلقی می‌شود.

3.داده مورد استفاده (نوع، منبع، اندازه)

نوع داده

- داده‌های متنی پست‌هایFacebook
- داده‌های تعاملی کاربران(Likes) ، Comments، (Shares
- داده‌های زمانی انتشار

منبع داده

- داده‌ها از مجموعه داده عمومی Facebook استخراج شده‌اند.
- منبع خام داده در GitHub و Zenodo منتشر شده است.

اندازه داده

- حدود **۳۵.۳۱ میلیون پستFacebook**
- بازه زمانی: فوریه ۲۰۲۱ تا ژانویه ۲۰۲۲
- حجم داده خام: حدود **۴۶.۵ گیگابایت**

- تعدادTopic های استخراج‌شده اولیه :**۱۰٬۹۳۲**
- پس از پاک‌سازی داده‌ها :**۳٬۵۷۰ Topic معتبر**
- توزیع کلاس‌ها:
  - ۱۵۷ Topic محبوب
  - ۳٬۴۱۳ Topic غیرمحبوب

عدم توازن شدید داده‌ها (۱ به ۲۲) یکی از چالش‌های اصلی مقاله است.

4.روش پیشنهادی مقاله (به زبان ساده)

گام 1: استخراجTopic ها

- پست‌های روزانه Facebook با استفاده از **TF-IDF**نمایش عددی می‌شوند.
- الگوریتم **K-Means**برای خوشه‌بندی متن‌ها استفاده می‌شود.
- هر خوشه به‌عنوان یک Topic در نظر گرفته می‌شود.
- Topicهای مشابه در روزهای مختلف با **Cosine Similarity**ادغام می‌شوند.

گام 2: محاسبه محبوبیتTopic

محبوبیت نهایی Topic به صورت زیر محاسبه می‌شود:

$$P = 0.46 \times Posts + 0.27 \times Shares + 0.17 \times Comments + 0.10 \times Likes$$

این وزن‌ها با ترکیب روش **AHP**و نظر خبرگان تعیین شده‌اند.

گام 3: استخراج ویژگی‌ها (۱۲ ساعت اول)

- استخراج ۳۸ ویژگی از متن، زمان و تعامل کاربران در ۱۲ ساعت اول انتشارTopic
- تحلیل اهمیت ویژگی‌ها نشان می‌دهد که:
  - Likes، Commentsو Shares اولیه بیشترین نقش را دارند.

گام 4: آماده‌سازی داده

- حذفTopic های کوتاه‌عمر یا ناقص

- رفع عدم توازن داده با:
  - **SMOTE** (Oversampling)
  - **Class Weighting**

گام 5: مدل‌سازی و یادگیری

مدل‌های زیر آموزش و مقایسه شده‌اند:

- Random Forest
- XGBoost
- LightGBM
- CatBoost
- SVM

شبه‌کد کلی روش

For each day:

Collect Facebook posts

Preprocess text

Cluster posts → Topics

For each Topic:

Extract first 12h features

Compute final popularity score

Assign label (Popular / Non-popular)

Balance dataset (SMOTE)

Train ML classifiers

Evaluate using Accuracy, Precision, Recall, F1, AUC

5.نتایج اصلی، محدودیت‌ها و ایده‌های ادامه

نتایج اصلی

- دقت کلی مدل‌ها بالاتر از ۹۰٪ است.
- Random Forestو XGBoost عملکرد متعادلی در Precision و Recall دارند.
- ویژگی‌های انتشار اولیه مهم‌ترین عامل پیش‌بینی هستند.
- روش پیشنهادی برای **پیش‌بینی زودهنگام** بسیار مناسب است.

محدودیت‌ها

- استفاده از داده‌های یک پلتفرم → (Facebook) تعمیم‌پذیری محدود
- عدم استفاده از مدل‌های عمیق مانند LSTM یاTransformer
- وابستگی به آستانه محبوبیت
- احتمال overfitting ناشی ازSMOTE

ایده‌های ادامه تحقیق

- استفاده از **Transformerها (BERT)** برای نمایش متنی قوی‌تر
- تحلیل کاربران اولیه منتشرکنندهTopic
- یادگیری وزن‌های محبوبیت به‌صورت داده‌محور
- گسترش مدل به **چند پلتفرمTwitter)** ، **Instagram**، **(...**
- ترکیب تشخیص ناهنجاری (Anomaly Detection) با طبقه‌بندی