

**بسم الله الرحمن الرحیم**

نام استاد:

دکتر نوا اسلامی

نام و نام خانوادگی گردآوران:

احسان لک،رضا کزازی

موضوع:

تحلیل احساسات

(Sentiment Analysis)

نام درس مربوطه:

پردازش زبان طبیعی

رشته و مقطع تحصیلی:

ارشد هوش مصنوعی و رباتیکز

پاییز 1403

چکیده

در این پژوهش، تحلیل احساسات بر اساس نظرات کاربران در فضای مجازی با استفاده از الگوریتم‌های یادگیری ماشین انجام شده است. هدف اصلی، طراحی مدلی برای طبقه‌بندی احساسات مثبت و منفی با دقت بالا بود. داده‌های دیجی کالا به عنوان ورودی مورد استفاده قرار گرفت و مراحل پیش‌پردازش شامل توکن‌سازی، حذف کلمات توقف و لماتیزه کردن اعمال شد. نتایج نشان داد که الگوریتم Naïve Bayes با دقت 89% درصد عملکرد بهتری نسبت به سایر الگوریتم‌ها داشت. این تحقیق پتانسیل بالای الگوریتم‌های ساده در تحلیل احساسات را نشان می‌دهد.

**مقدمه**

در این پروژه، هدف ما ساخت یک مدل تحلیل احساسات (Sentiment Analysis) برای متن‌های فارسی بود. تحلیل احساسات به عنوان یکی از زیرشاخه‌های پردازش زبان طبیعی (NLP)، به شناسایی و طبقه‌بندی احساسات موجود در متن‌ها می‌پردازد. در این کار، از الگوریتم‌های یادگیری ماشین مانند Naive Bayes و SVM استفاده شد. داده‌های مورد استفاده شامل کامنت‌های کاربران بود که به دو دسته‌ی مثبت و منفی تقسیم‌بندی شده‌اند.

**هدف**

هدف این پژوهش طراحی و توسعه مدلی است که بتواند احساسات مثبت و منفی را از نظرات کاربران با استفاده از الگوریتم‌های یادگیری ماشین تشخیص دهد و دقت بالایی در پیش‌بینی احساسات ارائه دهد.

روش شناسی

**1. مجموعه داده‌ها**

در این پژوهش از مجموعه داده‌های **دیجی کالا که در این ریپازیتوری استفاده شده [1]** [[1]](#footnote-1)استفاده شده است. این مجموعه شامل:

* زبان: فارسی
* تعداد کل داده ها: 2932747
* تعداد داده های استفاده شده: 2057167
* تعداد داده های مثبت : 1749055
* تعداد داده های منفی: 308112
* داده های اصلی استفاده شده پس از پیش پردازش 30000

**2. مراحل پیش‌پردازش**

۲. پیش‌پردازش داده‌ها:

برای آماده‌سازی داده‌ها و تبدیل آن‌ها به فرمتی که مدل‌های یادگیری ماشین بتوانند پردازش کنند، مراحل پیش‌پردازش زیر انجام شد:

۲.۱. نرمال‌سازی:

متن‌ها با استفاده از کتابخانه‌ی Hazm نرمال‌سازی شدند. این مرحله شامل تبدیل اعداد و حروف عربی به فارسی و یکسان‌سازی متن بود.

۲.۲. حذف لینک‌ها:

لینک‌های اینترنتی و تگ هایHTML (شروع‌شده با http، https یا www…,<p>) از متن حذف شدند. این کار با استفاده از عبارات باقاعده (regex) انجام شد.

۲.۳. حذف نام کاربری:

نام کاربری‌هایی که با @ شروع می‌شوند (مانند @username) از متن حذف شدند. این کار نیز با استفاده از عبارات باقاعده (regex) انجام شد.

۲.۴. تبدیل به حروف کوچک:

تمامی متن‌ها به حروف کوچک تبدیل شدند تا یکسان‌سازی شوند.

۲.۵. توکن‌سازی:

متن‌ها به کلمات کوچک‌تر (توکن‌ها) تقسیم شدند. این کار با استفاده از تابع word\_tokenize از کتابخانه‌ی Hazm انجام شد.

۲.۶. حذف کلمات توقف و علائم نگارشی:

کلمات توقف (مانند "و"، "در"، "است") و علائم نگارشی (مانند نقطه، ویرگول و ...) از متن حذف شدند. این کار با استفاده از لیست کلمات توقف (stopwords\_list) از کتابخانه‌ی Hazm انجام شد.

۲.۷. لماتیزه کردن(بن واژه):

کلمات به شکل پایه‌شان (ریشه) تبدیل شدند (مثلاً "رفتم" به "رفت"). این کار با استفاده از تابع Lemmatizer از کتابخانه‌ی Hazm انجام شد.

۲.۸. ترکیب توکن‌ها:

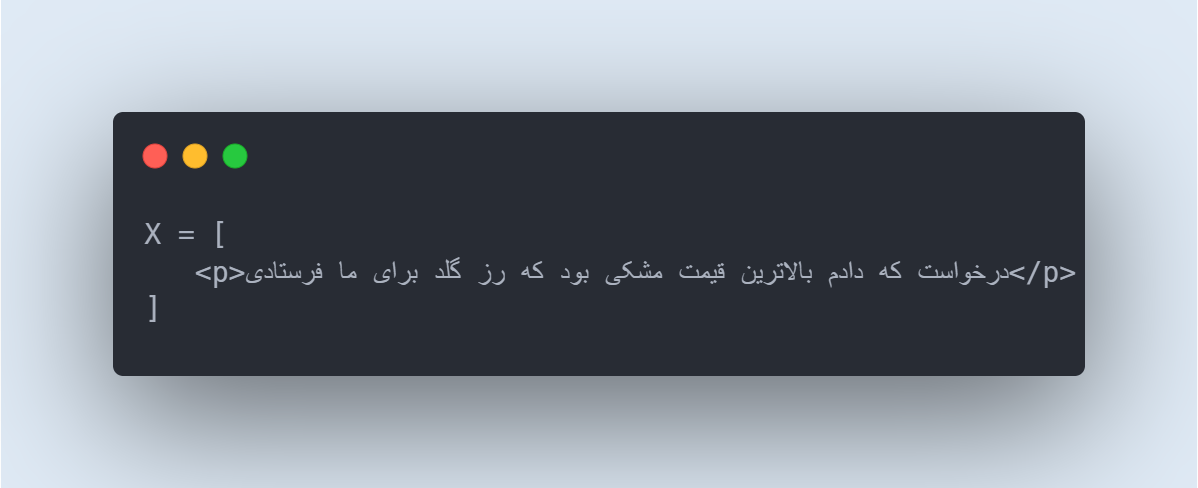
توکن‌ها به یک متن واحد تبدیل شدند تا برای پردازش بیشتر آماده شوند.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

۳. نمونه‌ای از داده‌های پردازش شده:

پس از انجام مراحل پیش‌پردازش، داده‌ها به شکل زیر تبدیل شدند:

داده‌های ورودی:

داده‌های پردازش شده:



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

۴. اهمیت پیش‌پردازش:

پیش‌پردازش داده‌ها یکی از مراحل کلیدی در پروژه‌های پردازش زبان طبیعی است. این مراحل به مدل کمک می‌کنند تا داده‌ها را به‌صورت دقیق‌تر و کارآمدتر پردازش کند. با حذف نویزها (مانند لینک‌ها، نام کاربری‌ها و کلمات توقف) و تبدیل کلمات به شکل پایه‌شان، دقت و عملکرد مدل بهبود می‌یابد

**3. مدل‌سازی و الگوریتم‌ها**

الگوریتم‌های مورد استفاده در این تحقیق:

* **Naïve Bayes:** یک الگوریتم ساده ولی قدرتمند بر اساس احتمالات.

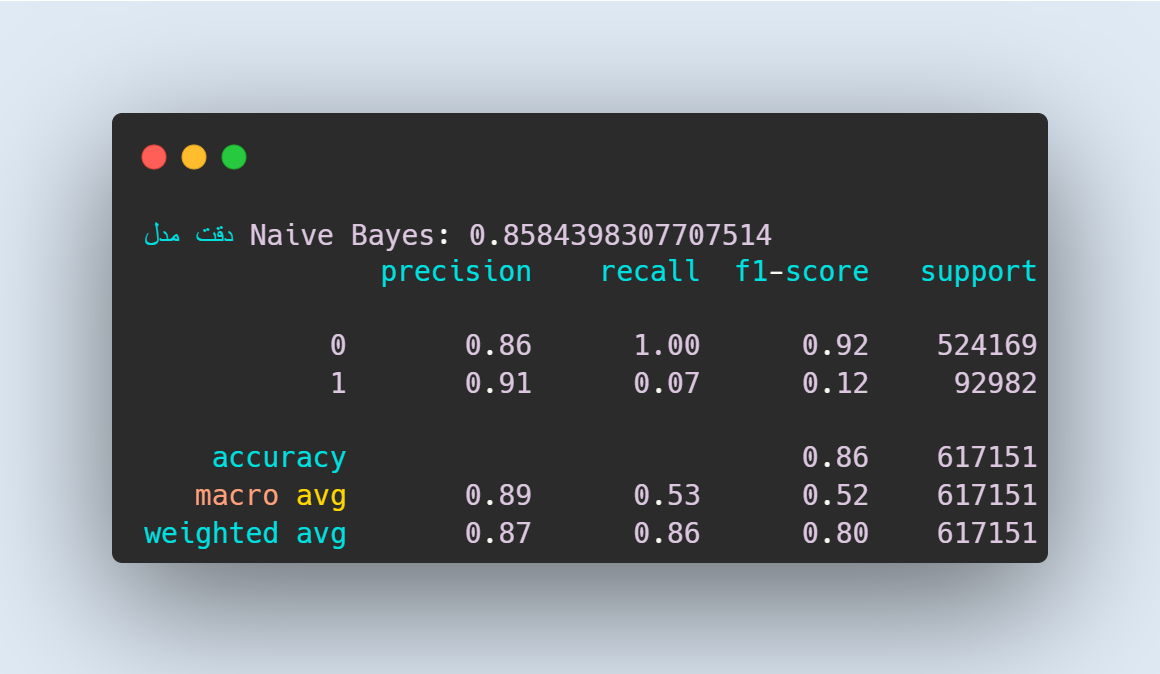
**4. تقسیم داده‌ها**

مجموعه داده‌ها به نسبت 70:30 به دو بخش تقسیم شدند:

* 70 درصد برای آموزش مدل‌ها.
* 30 درصد برای آزمایش عملکرد مدل‌ها.

نتایج و تحلیل

در این پژوهش دقت الگوریتم‌ در طبقه‌بندی داده به شرح زیر است:



نتایج نشان می‌دهند که الگوریتم Naïve Bayes با دقت بالایی را در این مجموعه داده‌ها ارائه داده است.

### نتیجه‌گیری

این پژوهش نشان داد که با استفاده از الگوریتم‌های ساده‌ای مانند Naïve Bayes می‌توان به دقت‌های بسیار بالایی در تحلیل احساسات دست یافت. با این حال، مدل‌ها در شناسایی احساسات پیچیده‌تر مانند کنایه عملکرد مطلوبی ندارند.

1. https://github.com/mosiomohsen/persian-sentiment-analysis-using-fastText-word-embedding-and-pseudo-labeling/blob/main/DK.part01.rar [↑](#footnote-ref-1)