JCWR2024
April 24-25, 2024; Tehran, Iran
10th International Conference on Web Research

LDPSA: A Large Dataset of Persian Sentiment Analysis

Ali Nazarizadeh*
Department of Computer Engineering, Central Tehran Branch, Islamic Azad University
Tehran, Iran
computer.engineer.as@gmail.com

Taha Samavati
Iran University of Science and Technology, Faculty of Computer Engineering
Tehran, Iran
taha_samavati@alumni.iust.ac.ir

Sina Mansouri
K. N. Toosi University of Technology, Faculty of Computer Engineering
Tehran, Iran
sina.mansouri79@email.kntu.ac.ir

Minoo Sayyadpour
Dept. Math and Computer Science
Kharazmi University
Tehran, Iran
std_minoosayyadpour@khu.ac.ir

***Abstract*—Today, data is more valuable to us than gold. When observing the environment, a substantial amount of data, particularly textual information, can be identified, tagged, prepared, and published in the form of a corpus or datasets. The primary objective of our paper is to gather, prepare, tag, and develop a vast dataset of Fidibo users' opinions regarding educational content and e-books. This dataset enables in-depth analysis of emotions and opinion mining, particularly within the educational content realm. A common flaw in nearly all similar datasets in the Farsi language is their restriction to user opinions on services and products available on online platforms. The dataset we refer to as LDPSA (A Large Dataset of Persian Sentiment Analysis) offers several advantages over comparable datasets in the Persian language. Notably, this dataset consists of 253,368 comments, each categorized into 5 classes. LDPSA represents the sole extensive Iranian dataset suitable for scrutinizing educational content and e-books. Moreover, significant insights were gleaned from data analysis. For example, during the COVID-19 pandemic, Iranian individuals dedicated more time to studying and engaging with educational platforms significantly. Nearly 80% of users expressed favorable opinions concerning the informational materials available on the Fidibo website. Users' inclination towards utilizing audio books has escalated, along with other analysis referenced in the paper.**

***Keywords*—Sentiment Analysis, Persian Sentiment Analysis, Corpus, Opinion Mining, Text Mining.**

JCWR2024
April 24-25, 2024; Tehran, Iran
10th International Conference on Web Research

LDPSA: مجموعه داده بزرگ آنالیز احساسات فارسی

علی نظری زاده[۱]، طاها سماواتی[۲]، سینا منصوری[۳]، مینو صیادپور[۴]

[۱] دپارتمان مهندسی کامپیوتر دانشگاه تهران مرکز
computer.engineer.as@gmail.com

[۲] گروه کامپیوتر دانشگاه ایران
taha_samavati@alumni.iust.ac.ir

[۳] گروه کامپیوتر دانشگاه خواجه نصیرالدین طوسی
sina.mansouri79@email.kntu.ac.ir

[۴] گروه ریاضی و کامپیوتر دانشگاه خوارزمی
std_minoosayyadpour@khu.ac.ir

چکیده

امروزه داده ها برای ما از طلا ارزشمندتر هستند. هنگام مشاهده محیط، مقدار قابل توجهی از داده ها، به ویژه اطلاعات متنی، را می توان شناسایی، برچسب گذاری، تهیه و در قالب مجموعه داده منتشر کرد. هدف اصلی مقاله ما جمع آوری، تهیه، برچسب گذاری و توسعه مجموعه وسیعی از نظرات کاربران فیدیبو در مورد محتوای آموزشی و کتاب های الکترونیکی است. این مجموعه داده تجزیه و تحلیل عمیق احساسات و عقیده کاوی، به ویژه در حوزه محتوای آموزشی را امکان پذیر می کند. یک نقص رایج تقریباً در تمام مجموعه داده های مشابه در زبان فارسی، محدودیت آنها به نظرات کاربران در مورد خدمات و محصولات موجود در پلتفرم های آنلاین است. مجموعه داده ای که ما از آن به عنوان LDPSA (مجموعه داده های بزرگ تحلیل احساسات فارسی) یاد می کنیم، مزایای متعددی نسبت به مجموعه داده های قابل مقایسه در زبان فارسی دارد. قابل ذکر است، این مجموعه داده شامل ۲۵۳۳۶۸ نظر است که هر کدام در ۵ کلاس طبقه بندی می شوند. LDPSA تنها مجموعه داده گسترده ایرانی است که برای بررسی دقیق محتوای آموزشی و کتاب های الکترونیکی مناسب است. علاوه بر این، بینش های قابل توجهی از تجزیه و تحلیل داده ها به دست آمد. به عنوان مثال، در طول همه‌گیری COVID-19، افراد ایرانی زمان بیشتری را به مطالعه و تعامل با بسترهای آموزشی اختصاص دادند. نزدیک به ۸۰ درصد از کاربران نظر مساعد خود را در مورد مطالب اطلاع رسانی موجود در وب سایت فیدیبو اعلام کردند. تمایل کاربران به استفاده از کتاب های صوتی، همراه با تحلیل های دیگری که در مقاله به آنها اشاره شده است، افزایش یافته است.

کلمات کلیدی: آنالیز احساسات، آنالیز احساسات فارسی، پیکره متنی، نظر کاوی، پردازش متن.