Twitter өгөгдөлд суурилсан сэтгэл хөдлөлийг тодорхойлох нь

1

2025 оны 11-р сарын 22

Энэхүү төслийн ажлын хүрээнд нийгмийн сүлжээний Sentiment140 өгөгдлийн санд машин сургалтын арга зүйг ашиглан хэрэглэгчдийн сэтгэл хөдлөлийг эерэг эсвэл сөрөг гэж ангилах загварыг хөгжүүлэв. Судалгаанд Python хэлний Scikit-learn санг ашиглан өгөгдлийг цэвэрлэж, CountVectorizer болон TF-IDF аргуудаар текстэн өгөгдлийг тоон хэлбэрт шилжүүлсэн. Сэтгэл хөдлөлийг таамаглахын тулд Мультиномиал Наив Байес (Multinomial Naive Bayes) болон Ложистик Регресс (Logistic Regression) алгоритмуудыг харьцуулан сургав. Үр дүнг Streamlit дээрх интерактив самбар (dashboard) ашиглан магадлалын тархалт болон төөрөгдлийн матрицаар дүрслэн харуулав. Туршилтын үр дүнд Ложистик Регресс/Мультиномиал Наив Байес загвар нь [X]%-ийн нарийвчлалтайгаар хамгийн өндөр үр дүнг үзүүлсэн байна.

Агуулга

Шаардлагатай багцууд

Шаардлагатай багцуудыг дараах байдлаар урьдчилан суулгана.

```
pip install streamlit pandas numpy scikit-learn matplotlib seaborn plotly
```

1 Өгөгдөл

Энэхүү судалгааны ажилд бид нийгмийн сүлжээний сэтгэл хөдлөлийг шинжлэхэд өргөн хэрэглэгддэг Sentiment140 [1] өгөгдлийн санг ашиглав. Тус өгөгдлийн сан нь Twitter (одоогийн X) сүлжээнээс цуглуулсан жиргээнүүдээс бүрдэх бөгөөд хэрэглэгчийн сэтгэл хөдлөлийг “Эерэг” (Positive) болон “Сөрөг” (Negative) гэж урьдчилан ангилсан байдаг.

1.1 Өгөгдлийн түүвэрлэлт ба Боловсруулалт

Sentiment140 өгөгдлийн сан нь нийт 1.6 сая мөр бүхий их хэмжээний өгөгдлийг агуулдаг. Энэхүү төслийн хүрээнд тооцооллын нөөц болон хугацааг хэмнэх зорилгоор бид “Санамсаргүй түүвэрлэлт” (Simple Random Sampling)-ийн аргыг ашиглав. Бид Python хэлний sample() функцийг ашиглан нийт өгөгдлөөс 50,000 жиргээг санамсаргүй байдлаар сонгон авч сургалтад ашигласан. Туршилтын үр дүн дахин давтагдах (reproducible) боломжтой байхын тулд random_state=42 тохиргоог ашигласан болно.

```
# Бүтэн өгөгдлийг унших (1.6 сая мөр)
df_full = pd.read_csv("training.1600000.processed.noemoticon.csv",
                      encoding="latin-1", header=None)

# 50,000 мөрийг санамсаргүйгээр түүвэрлэх (Seed = 42)
df_sample = df_full.sample(n=50000, random_state=42)
```

Дүгнэлт

Ашигласан материал

[1] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision.” Stanford, 2009.