Processament de Llenguatge Natural aplicat a ressenyes de videojocs.

Víctor Fernández Florensa

Grau en Enginyeria Informàtica

Treball dirigit per: Nil Torrent Bureu i Jordi Planes Cid
Departament: Enginyeria Informàtica i Disseny Digital
Campus de Cappont. Edifici EPS. Despatx 2.17
656434136, victorfflorensa@gmail.com

Resum

El present treball se centra en el Processament del Llenguatge Natural (PLN) i l'anàlisi de sentiments per analitzar les ressenyes d'una saga de videojocs molt coneguda: "The Last of Us".

En el treball s'han elaborat diverses hipòtesis sobre els conjunts de dades i, mitjançant l'anàlisi de dades, s'han intentat respondre. També, s'han aplicat tècniques de preprocessament de text i s'han utilitzat diversos models Transformadors per manipular les dades o generar-ne de noves.

1. Introducció

Els motius pels quals es va decidir fer aquest treball van ser primerament, per continuar aprenent en el camp de la ciència de dades i la intel·ligència artificial, i, per altra banda, per relacionar-ho amb una de les meves sagues de videojocs preferides com és "The Last of Us".

L'interès pel conjunt de dades escollit també es va deure a la polarització que va generar la preqüela The Last of Us: Part II, que va dividir l'opinió dels usuaris i de la premsa aportant ressenyes molt diverses.

Per realitzar aquest treball es va dur a terme una feina de documentació sobre el context del problema i una introducció al Processament de Llenguatge Natural i l'anàlisi de sentiments.

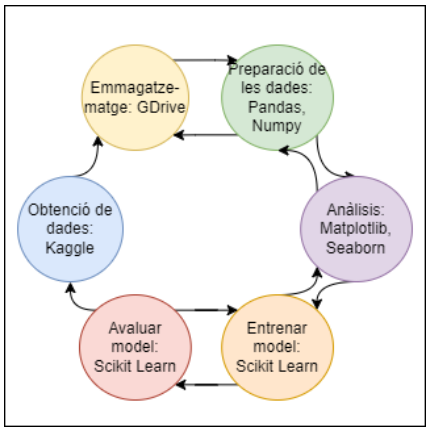
2. Desenvolupament del treball

La metodologia utilitzada per l'elaboració de la part pràctica ha estat CRISP-DM, la qual estructura el projecte en diverses fases des de la comprensió inicial del problema fins a l'anàlisi i l'avaluació dels resultats.

En la fase inicial, es va definir els objectius del TFG i es va dur a terme una anàlisi preliminar de les dades disponibles.

Aquest procés incloïa una exploració inicial per identificar possibles problemes o mancances en les dades i una elaboració d'hipòtesis inicials.

Posteriorment, es va procedir a la preparació de les dades, que incloïa la neteja i una exploració visual per detectar correlacions i distribucions de les dades.



Il·lustració 1 Diagrama del cicle de vida de les dades

Un cop les dades estaven preparades, es van aplicar diversos models de PLN per analitzar els sentiments expressats en les ressenyes. Es van utilitzar models preentrenats com twitter-xlm-roberta-base-sentiment, sentiment-roberta-large-english, SaBERT-Spanish-Sentiment-Analysis, rubert-tiny2-russian-sentiment, i el model Vader. Això va permetre poder classificar les dades segons la seva polaritat, positiva, negativa o neutra i explorar patrons ocults en les dades. Per exemple, els comentaris classificats com negatius amb puntuacions de 10 i viceversa. També va servir per veure una evolució del sentiment al llarg del temps i confirmar algunes hipòtesis.

Addicionalment es va elaborar un model de freqüència de paraules (Bag of Words) i una representació gràfica (Word Cloud) que va suposar haver de traduir tot el conjunt de dades per tenir uns resultats els més objectius possibles. Es van explorar dos models de traducció: el model de Meta, SeamlessM4T-v2, en la seva versió de text a text; i el model multilingüe de l'Open Parallel Corpus, opus-mt-mul-en. A més, es va fer una feina de preprocessament de les dades, utilitzant tècniques com la tokenització, lemmatització, StopWord Removal, NER (Name Entity Recognition), entre altres.

A més a més, es va entrenar un model de regressió per, a partir dels valors de polaritat de l'anàlisi de sentiments, fer

un model per predir de la puntuació dels usuaris. Finalment, també es va explicar totes les mètriques utilitzades per avaluar el seu rendiment.

A continuació s'adjunta una taula amb els resultats obtinguts.

Mètrica	Resultats Regressió Lineal amb validació creuada k=5
MAE (Mean Absolute Error)	1.499
MSE (Mean Squared Error)	4.994
RMSE (Root Mean Squared Error)	2.234
R-Squared	0.742

3. Conclusions

En el treball de final de grau s'ha realitzat una anàlisi exploratòria del conjunt de dades que ha comportat la investigació d'algunes hipòtesis.

- S'ha desmentit que l'opinió del primer videojoc millorés amb el pas del temps amb la sortida del segon.
- S'ha trobat una correlació directa entre un increment en la quantitat de ressenyes i notorietat del primer videojoc amb la sortida del segon, a causa de la continuïtat directa de la història.
- S'ha trobat diverses evidències que confirmen que el factor determinant en l'opinió polaritzada dels usuaris respecte a la segona part és degut principalment a les decisions de la narrativa.
- També que els usuaris tendeixen a valorar més la narrativa per damunt de l'apartat tècnic.
- S'ha analitzat altres variables d'interès del conjunt de dades com els vots i les visualitzacions dels comentaris a la pàgina de Metacritic i s'ha observat que estan estretament relacionats amb l'opinió general del joc.
- S'ha corroborat que l'opinió del segon videojoc sí que ha millorat amb el pas dels temps.

A més, també s'ha complert l'objectiu d'entrenar un model que predigués les notes dels usuaris a partir de les ressenyes dels usuaris.