

ΘΕΜΑ: Ανάλυση Συναισθήματος σε Ελληνικά & Αγγλικά Κείμενα με Χρήση LLMs

Στόχος της εργασίας είναι η συγκριτική αξιολόγηση σύγχρονων LLMs (π.χ. GPT-4, Gemini) στην ανάλυση συναισθήματος ελληνικών και αγγλικών κειμένων, σε σενάρια zero-shot και few-shot, με έμφαση σε περιπτώσεις όπως ειρωνεία, αμφισημία ή και code-switching (εναλλαγή δύο γλωσσών μέσα σε μια πρόταση).

A. ΕΠΙΣΚΟΠΗΣΗ ΤΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

ΒΗΜΑ 1 (MEXPI 4/11):

- Αναζήτηση βιβλιογραφίας στο scholar.google.com με λέξεις, προτάσεις-κλειδιά όπως:
 - “sentiment analysis with large language models”
 - “zero-shot sentiment classification”
 - “Greek sentiment analysis”
- Παρουσίαση υπαρχόντων μελετών:
 - Παραδοσιακά μοντέλα (SVMs, RF, LSTMs).
 - Προηγούμενες εργασίες με BERT/Transformers.
 - Νεότερες προσεγγίσεις με GPT, Gemini.

B. ΕΠΙΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

ΒΗΜΑ 2 (MEXPI 7/11):

Επιλογή dataset σε δύο γλώσσες (EN/EL) για την ανάλυση συναισθήματος.

Μερικά παραδείγματα (Αγγλικά Datasets):

1. IMDB Movie Reviews Dataset

Πηγή: Kaggle

(<https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>) ή TensorFlow Datasets

2. Amazon Product Reviews Dataset

Πηγή: Amazon Review Data (Ni et al., 2019)

(<https://nijianmo.github.io/amazon/index.html>)

3. **Sentiment140 (Twitter Dataset)**

Πηγή: Huggingface

(https://huggingface.co/datasets/stanfordnlp/sentiment140?utm_source=chatgpt.com)

4. **Yelp Reviews Dataset**

Πηγή: Yelp Open Dataset (<https://business.yelp.com/data/resources/open-dataset/>)

Σχέδιο Χρήσης Δεδομένων

- Μετάφραση του αγγλικού κειμένου στα ελληνικά με τη χρήση αυτόματου συστήματος μετάφρασης.
- Επιλογή δείγματος ~500–1000 παραδειγμάτων, με ισορροπία σε θετικά, αρνητικά ή και ουδέτερα παραδείγματα (ανάλογα με το dataset).
- Προαιρετικά: Εστίαση και σε ειδικές περιπτώσεις, όπως ειρωνεία ή σαρκασμός.

Γ. ΕΦΑΡΜΟΓΗ ΣΥΣΤΗΜΑΤΩΝ

ΒΗΜΑ 3 (MEXPI 28/11):

- Επιλογή LLMs: πχ GPT-4, Gemini.
- Ορισμός πειραματικών σεναρίων:
Θα τρέξετε τα ίδια κείμενα σε δύο στησίματα: χωρίς παραδείγματα (zero-shot) και με λίγα ενδεικτικά παραδείγματα (few-shot). Οι ετικέτες που ζητάμε από το μοντέλο είναι σταθερές σε όλα τα runs (π.χ. Θετικό/Αρνητικό/Ουδέτερο).
- **Zero-shot:**
Δώστε στο μοντέλο μόνο το κείμενο και σαφείς οδηγίες για την ετικέτα που πρέπει να επιστρέψει. Δεν βάζουμε κανένα παράδειγμα λύσης μέσα στο prompt. Ζητάμε σύντομη και καθαρή απάντηση (π.χ. μόνο την ετικέτα και μία σύντομη αιτιολόγηση).
- **Few-shot (3–5 παραδείγματα):**
Πριν από το κείμενο προς αξιολόγηση, δίνουμε 3–5 μικρά, αντιπροσωπευτικά παραδείγματα με τη σωστή ετικέτα. Τα παραδείγματα δεν πρέπει να προέρχονται από τα τεστ κείμενα. Φροντίζουμε να καλύπτουν τυπικές περιπτώσεις (π.χ. θετικό, αρνητικό, ουδέτερο).
- Συλλογή αποτελεσμάτων για κάθε dataset.

Δ. ΑΞΙΟΛΟΓΗΣΗ

ΒΗΜΑ 4 (MEXPI 9/12):

- Αυτόματη αξιολόγηση με μετρικές:
 - Accuracy, Precision, Recall, F1-score.
 - Confusion matrix (για να δούμε πού μπερδεύεται το μοντέλο).

Με τη χρήση της ιστοσελίδας <https://mateo.ivdnt.org/Translate>

- Ποιοτική ανάλυση:
 - Περιπτώσεις ειρωνείας/σαρκασμού.
 - Πολύ μεγάλες ή ασαφείς προτάσεις.
 - Παραδείγματα με code-switching (μείξη EL/EN).

Ε. ΣΥΓΚΡΙΣΗ ΣΥΣΤΗΜΑΤΩΝ

ΒΗΜΑ 5 (ΜΕΧΡΙ 16/1):

- Πίνακας με συγκριτικά αποτελέσματα GPT-4, Gemini, ίσως και άλλα.
- Ανάλυση διαφορών:
 - Ποιο είναι πιο “συντηρητικό” (π.χ. δίνει πιο πολλές ουδέτερες).
 - Ποιο είναι πιο ευαίσθητο σε σαρκασμό.

ΣΤ. ΔΟΜΗ ΕΡΓΑΣΙΑΣ

ΒΗΜΑ 6 (ΜΕΧΡΙ ΤΗΝ ΗΜΕΡΟΜΗΝΙΑ ΤΩΝ ΕΞΕΤΑΣΕΩΝ ΤΟΥ ΜΑΘΗΜΑΤΟΣ):

Συντάξτε επιστημονική εργασία στα Αγγλικά (σαν τις εργασίες της βιβλιογραφίας) μέχρι 10 σελίδες το πολύ.

Η εργασία θα περιλαμβάνει:

1. Θεωρητικό υπόβαθρο (sentiment analysis + LLMs).
2. Περιγραφή dataset (με στατιστικά).
3. Πειραματική διαδικασία (zero-shot/few-shot).

4. Αποτελέσματα & γραφήματα (accuracy, F1).
5. Ανάλυση σφαλμάτων.
6. Συμπεράσματα: τι δουλεύει καλύτερα, τι όχι.
7. Μελλοντικές βελτιώσεις.
8. Βιβλιογραφία.

ΑΛΛΑ

Η εργασία θα πραγματοποιηθεί σε ομάδες των τεσσάρων τουλάχιστον ατόμων.

Η εργασία είναι υποχρεωτική και πιάνει το 50% στον τελικό βαθμό.

Η ομάδα που θα πετύχει το μεγαλύτερο σκορ στα αποτελέσματα θα έχει

- επιπλέον βαθμολογικό bonus
- ευκαιρία να προωθήσει την εργασία της για δημοσίευση σε επιστημονικό συνέδριο