Sentiment Analysis and Opinion Detection with Advanced Machine Learning Techniques

UNIVERSITY OF ŽILINA
Faculty of Management Science and Informatics

Ing. Peter Szathmáry   **Supervisor: Ing. Lukáš Falát, PhD.**

Motivation

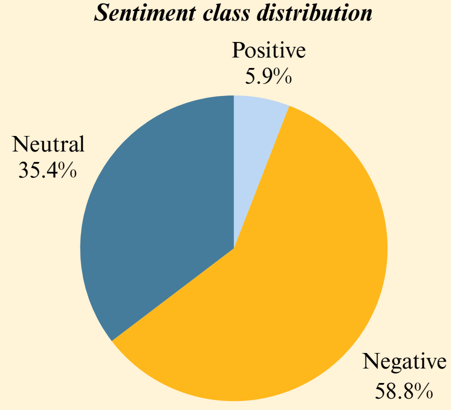
Analyzing opinions and sentiment in text is crucial for marketing, politics, and customer support. The main challenges are the ***limited amount of labeled data, imbalanced classes, and the need for higher accuracy.*** This work explores modern methods to overcome these challenges and improve model reliability.

Goals

- Design and implement models for ***sentiment analysis*** and ***opinion detection***.
- Test the effect of four improvement techniques:

embedding fusion | *Focal Loss* | *Standard Data Distillation (SDD)* | *Filtered Data Distillation (FDD)*

- Evaluate results with a ***systematic experimental pipeline***, including GridSearch hyperparameters tuning and statistical testing.



Methodology

The proposed framework:

1.DATA PROCESSING

- text cleaning
- tokenization
- lemmatization
- embeddings

2.MODELING

Two deep neural networks (for sentiment and opinion detection). Unlike **transformer models**, our approach uses **lightweight neural networks**, requiring **less hardware** and enabling deployment on systems with limited computational capacity.

3.ENHACEMENT TECHNIQUES:

- **Embedding fusion** (GloVe, Word2Vec, FastText)
- **Focal Loss for class imbalance**
- **Standard Data Distillation** (SDD)
- **Filtered Data Distillation** (FDD) – **novel method** introduced in this thesis as an enhancement of SDD, designed to reduce noise in pseudo-labels

4.EVALUATION

- 5-fold cross-validation
- 10 replications
- **Mann–Whitney U test**

The dataset consisted of ***7 830 posts*** from the Bluesky platform, with ***800 manually annotated samples*** (sentiment + presence of opinion).

In total, ***2 425 models*** were trained and evaluated.

Results

Model/Experiment	Sentiment F1	Opinion F1	Statistical significance
Baseline	59.97%	62.20%	-
RQ1: Embedding fusion	**62.88%**	**64.89%**	✓ ($p = 0.0002$ / 0.0445)
RQ2: Focal Loss	60.18%	-	✗ ($p = 0.3957$)
RQ3: Standard DD (SDD)	61.96%	62.73%	✓ for sentiment only
RQ4: Filtered DD (FDD)	**62.59%**	**64.89%**	✓ ($p = 0.0002$ / 0.0188)

- ***Best results:*** Embedding fusion (RQ1) and **FDD (RQ4)**.
- ***Focal Loss*** did not bring significant improvement.
- ***FDD (novel method)*** achieved the strongest gain *in opinion detection.*

Conclusion

- ***Most effective techniques***: Embedding fusion and ***Filtered Data Distillation (FDD).***
- ***FDD is an original contribution of this thesis***, not previously published, and it improves on Standard Data Distillation (SDD) by filtering out uncertain pseudo-labels.
- Data Distillation methods help extend training data when annotations are limited.
- Unlike resource-intensive transformer models, this approach is ***lightweight, practical, and suitable for deployment*** in environments with limited hardware.
- Results show potential for **applications** in analyzing customer feedback, monitoring public opinion, and evaluating product reviews.

Future work

- Explore transformer-based embeddings (BERT, RoBERTa).
- Address class imbalance with oversampling/undersampling or ensembles.
- Improve annotation quality via multiple annotators and inter-rater agreement.