Multi-source Multi-domain Sentiment Analysis with BERT-based Models

Gabriel Roccaruna, *Steve Azzolin, Giuseppe Riccardi*

Signals and Interactive Systems Lab, DISI, University of Trento

Outline

1. Introduction

Sentiment Analysis

- Extract the writer's attitude from written texts.

"The first Star Wars movie was terrific!" - ***Positive***

BERT-based models

- Achieved *state-of-the-art* results
- **Missing:** evaluation of performance variability on Multi-source Multi-domain
  - **Source:** where documents come from (Amazon, Twitter, IMDB, etc.)
  - **Domain:** category of the documents (travel, financial, health, etc.)

2. Our Contributions

Performance Variability Analysis

- We computed the performance variability of a BERT-based model on **eight Italian corpora** of different domains and sources

Improved Sentiment Analyzer based on BERT

- We optimized and improved a pre-existing BERT model for the Sentiment Analysis task

Novel publicly available dataset

- We collected and annotated a novel financial dataset in the Italian language.

3. Models

Baselines

- Prior-Label-Distribution classifier
- OpeNER (lexicon-based model)

BERT-based

- FEEL-IT (fine-tuned on FEEL-IT, predicts only positive and negative)
- AlBERTo Multi-Class (AMC) (fine-tune on Sentipolc16)
  - AlBERTo Multi-Class optimized (AMC opt)

4. Copora

Name	Source	Domain	Public Available	Collected By Us
SENTIPOLC16	Twitter	Socialpolitcal	✓	x
FEEL-IT	Twitter	General	✓	x
MultiEmotions-IT	Youtube&Facebook	Comments	✓	x
Trip-MAML	Tripadvisor	Hotel	✓	x
AriEmozione 1.0	Opera	Opera	✓	x
COADAPT	Personal Narratives	Psychology	x	✓
Italian Twitter Financial News	Twitter	Financial	✓	✓
Amazon Reviews	Amazon	Products	x	✓

5. Italian Twitter Fianancial News (ITFN)

ITFN

- Collection of 6040 tweets from financial journals
  - *Milano Finanza, Wall Street Italia, Il Sole 24 Ore* and *Italia Oggi*
- Random subsample from a pool of 84562 tweets written over 3 years (2018,2019,2020)

Preprocessing

- Automatic splitting of tweets into Functional Units (FU)
- FU is the minimal span of text with a communicative function
  - concept borrowed from the Dialogue Act theory (ISO 24617-2)
- 8529 FU produced

Positive

Industry, evidence of recovery also in July. But the downward trend is 8%

Negative

5.1 ITFN Annotation

Annotation Task

- 3 voluntarily annotators
- Identify the sentiment of a news by looking at the writer's intention.
  - *Writing style and terminology used*
- The labels were *positive, negative, neutral and Not Applicable (NA)*
  - *NA used to mark the spam and errors in the segmentation (640 FUs)*

# Examples	7889
Positive	34 %
Negative	31 %
Neutral	35 %

Tab. 1 Label distribution of ITFN

6. The role of neutrals

- **Inter-Annotator Agreement (IAA):**
  - Krippendorf's Alpha is 0.54 (1.0 is full agreement)
- **Impacts of neutrals on IAA**
  - **Removing** examples with at least one neutral IAA increases substantially
    - *We observed this in ITFN, COADAPT and MultiEmotions-IT*
- **Segmentation into Functional Units**
  - The agreement on tweets with more than one functional unit is higher than tweets with one functional unit *(0.56 vs 0.51)*

Labels	ITFN	COADAPT
-2,-1, 0,1,2	–	0.67
neg, neu, pos	0.54	0.73
neg, pos	0.94	0.98

Fig. 1 IAA computed with Krippendorf's Alpha

7. Experiments

1) AMC opt evaluation

- AMC opt vs FEEL-IT, AMC and Baselines
- AMC opt performance variability

2) single-source single-domain (ss-sd)

- Fine-tuning and testing on the same corpus
- Measure the maximum performance

3) multi-source multi-domain (ms-md)

- Fine-tuning and testing on a joint training and test sets
- Investigate the universal usage of the model

8. Results

Corpus Source Domain # Test	**SENTIPOLC16** Socio-political 1964	**ITFN** Twitter Financial 785	**FEEL-IT** General 2037	**MultiE.** YT/FB Comments 486	**Amazon** Amazon Products 125	**Trip-M.** Tripadvisor Hotels 352	**Ari** Opera Opera 250	**COADAPT** PHA Psychology 439
PLD	0.34/ 0.50*	0.37/ 0.51*	—	0.37/ 0.49*	0.35/ 0.52*	0.32/ 0.56*	0.30/ 0.45*	0.30/ 0.52*
OpeNER	0.28/ 0.40*	0.40/ 0.59*	0.59	0.40/ 0.61*	0.35/ 0.56*	0.50/ 0.74*	0.33/ 0.60*	0.60/ 0.64*
AMC	0.64/ 0.76*	0.41/ 0.58*	**0.89**	0.66/ 0.82*	0.44/ 0.71*	0.48/ 0.79*	0.42/ 0.61*	0.64/ 0.88*
FEEL-IT	0.39/ 0.84*	0.31/ 0.57*	—	0.48/ 0.76*	0.50/ 0.82*	0.60/ 0.91*	0.40/ 0.62*	0.31/ 0.81*
AMC opt	**0.69/ 0.82***	0.44/ 0.73*	0.87	0.68/ 0.85*	0.51/ 0.78*	0.55/ 0.88*	0.45/ 0.66*	0.64/ 0.82*
AMC opt ss-sd	—	**0.66/ 0.85***	—	0.73/ 0.87*	**0.65/ 0.88***	0.58/ 0.91*	0.73/ 0.74*	0.76/ 0.90*
AMC opt ms-md	0.62/ 0.82*	0.64/ 0.84*	0.87	**0.74/ 0.89***	0.63/ 0.88*	**0.67/ 0.94***	**0.75/ 0.76***	**0.77/ 0.92***

Tab 1. Macro-F1results for the three experiments. ***evaluation only on positives and negatives**

ss–sd = single-source single-domain

ms–md = multi-source multi-domain

8. Results

Corpus Source Domain # Test	**SENTIPOLC16** Socio-political 1964	**ITFN** Twitter Financial 785	**FEEL-IT** General 2037	**MultiE.** YT/FB Comments 486	**Amazon** Amazon Products 125	**Trip-M.** Tripadvisor Hotels 352	**Ari** Opera Opera 250	**COADAPT** PHA Psychology 439
PLD	0.34/ 0.50*	0.37/ 0.51*	—	0.37/ 0.49*	0.35/ 0.52*	0.32/ 0.56*	0.30/ 0.45*	0.30/ 0.52*
OpeNER	0.28/ 0.40*	0.40/ 0.59*	0.59	0.40/ 0.61*	0.35/ 0.56*	0.50/ 0.74*	0.33/ 0.60*	0.60/ 0.64*
AMC	0.64/ 0.76*	0.41/ 0.58*	**0.89**	0.66/ 0.82*	0.44/ 0.71*	0.48/ 0.79*	0.42/ 0.61*	0.64/ 0.88*
FEEL-IT	0.39/ 0.84*	0.31/ 0.57*	—	0.48/ 0.76*	0.50/ 0.82*	0.60/ 0.91*	0.40/ 0.62*	0.31/ 0.81*
AMC opt	**0.69/ 0.82***	0.44/ 0.73*	0.87	0.68/ 0.85*	0.51/ 0.78*	0.55/ 0.88*	0.45/ 0.66*	0.64/ 0.82*
AMC opt ss-sd	—	**0.66/ 0.85***	—	0.73/ 0.87*	**0.65/ 0.88***	0.58/ 0.91*	0.73/ 0.74*	0.76/ 0.90*
AMC opt ms-md	0.62/ 0.82*	0.64/ 0.84*	0.87	**0.74/ 0.89***	0.63/ 0.88*	**0.67/ 0.94***	**0.75/ 0.76***	**0.77/ 0.92***

Tab 1. Macro-F1results for the three experiments. ***evaluation only on positives and negatives**

ss–sd = single-source single-domain

ms–md = multi-source multi-domain

9. Error Anlysis

Inspection of F1-score distributions of AMC opt and AMC opt ss-sd

- **Positive and negative classes**
  - Consistently captured
  - Evident benefit of fine-tuning

- **Neutral class**
  - Large variance across corpora
  - Lowest median score

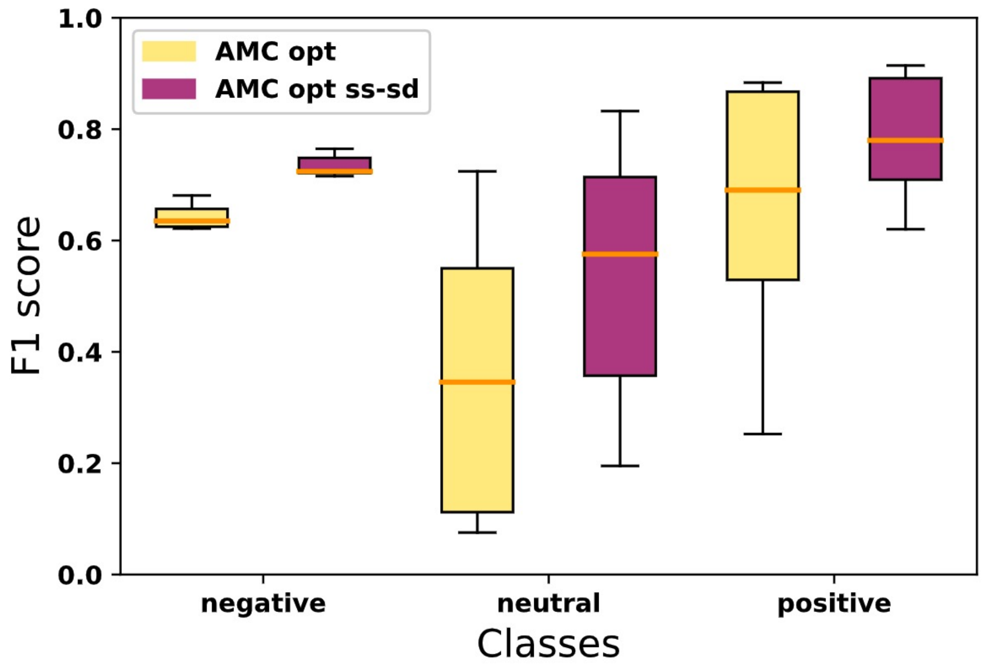

Fig 2. *Per-class F1-score distribution across corpora for the AMC opt model and the AMC opt model fine-tuned with the single-source single-domain setup.*

10. Conclusions

- **A novel Italian Finacial dataset annotated with sentiment (ITFN)**
  - Publicly available
- **BERT-based models attain robust performance variability**
  - Universal applicability across different sources and domains
- **Neutral class issues**
  - Humans and Machine have difficulty in the recognition
  - Ambiguities could be mitigated by segmenting the text into functional units

THANK YOU