

Use case 1 – NLP a analýza recenzí

Dostupné podklady

- Stručný popis datasetu: textové recenze + label pos/neg.
- Ukázka několika recenzí (2 pozitivní, 2 negativní, 1 ironická / sarkastická).
- Tabulka výsledků dvou textových reprezentací (BoW vs TF-IDF) + jednoho klasifikátoru.
- Pár příkladů chybné klasifikace (recenze a predikovaná vs. skutečná třída).

Reprezentace Accuracy F1 (pos) F1 (neg)

BoW	0.82	0.80	0.84
TF-IDF	0.86	0.85	0.87

Otázky

1. Rozdíl BoW vs. TF-IDF

Na základě Tabulky vysvětlete:

- a) V čem se konceptuálně liší přístup **Bag-of-Words** a **TF-IDF**.
- b) Proč může TF-IDF dosahovat lepších výsledků u recenzí, kde se některá slova vyskytují velmi často (např. “good”, “product”).
- c) Uveďte konkrétní typ slov, u kterých byste očekávali, že TF-IDF jim sníží váhu, a proč.

2. Stop slova – kdy (ne)odstraňovat

Představte si, že model špatně rozlišuje mezi větami:

- “I am **not** happy with this product.”
- “I am happy with this product.”

- a) Vysvětlete, jak může **agresivní odstranění stop slov** zhoršit kvalitu modelu.
- b) Uveďte **dvě situace**, kdy je odstranění stop slov naopak užitečné.
- c) Navrhněte, jak byste přistoupili k práci se stop slovy v **krátkých recenzích**.

3. Omezení klasických přístupů a návrh zlepšení

Vysvětlete, proč přístupy založené na BoW/TF-IDF:

- nerozumí pořadí slov,
- neumí dobře zachytit sarkasmus,
- mají problém se synonymy.

Navrhněte **dva konkrétní modernější přístupy** a stručně popište, v čem by mohly být pro analýzu recenzí lepší. Není třeba detailní technický popis, stačí intuitivní vysvětlení.