Рецензия на статью "Выявление поляризации и нейтральности текстов в новостном потоке"

Рецензент: **Абрамов В. А.**

Резюме

В статье предлагается метод выделения поляризаций текстовых документов, основанный на трёхступенчатом подходе:

1. Выделение нерелевантных документов на основе семантической близости и моделей OneClassSVM или IsolationForest
2. Выделение нейтральных документов на основе баланса токенов положительной и отрицательной окраски
3. Применение Elbow technique для определения числа кластеров

Алгоритм тестируется на одном датасете и по качеству разметки близок к человеческой.

Достоинства

1. Исследование описано очень подробно, все подбираемые параметры и выбор моделей обоснованы
2. Имеются графики, отображающие зависимости метрик от параметров
3. Имеются графики, отображающие характеристики данных из датасета
4. Подробно описана изучаемая задача
5. Предложенное решение почти не уступает человеческой разметке
6. Используются классические алгоритмы, нет лишней сложности

Критика

1. Не хватает обзора литературы по теме исследования, ссылок на другие статьи мало
2. Описание данных следует сократить, таблички на страницах 3-5 можно отправить в приложение. Суть можно свести к паре предложений о нейтральных и нерелевантных сообщениях в корпусах
3. Структура статьи похожа на отчет о проделанной работе - хотелось бы видеть (до экспериментов) описание похожих методов, краткое описание нового метода и потом эксперименты

4. Некоторые эксперименты (сравнение способов подсчета метрик, сравнение K-Means и K-Means++) являются техническими нюансами, это можно также вынести в приложение в ablation study
5. Нет описания алгоритма в форме псевдокода
6. Некоторые графики трудно читать (см. reachability chart стр. 8)
7. Страница 10 (Модифицированный алгоритм):
    - *"Как было сказано в начале текущего раздела, мы не используем значения семантической близости всех пар документов"* - если я правильно понимаю, то значения близости всё же используются, но вместо всех объектов выделяется подмножество ближайших
    - *"Хотя на начальных этапах был вариант алгоритма, где для каждого документа считалась сумма расстояний до всех остальных сообщений"* - можно это также вынести в приложение или провести сравнение предлагаемой модели и этого подхода в экспериментах
8. В выводах можно дописать своё мнение о плюсах и минусах полученной модели, добавить абзац про дальнейшее направление исследований