Products opinions and news
Project Proposal for NLP Course, Winter 2023

Anish Gupta
Warsaw University of Technology
01175535@pw.edu.pl

Martyna Majchrzak
Warsaw University of Technology
martyna.majchrzak.stud@pw.edu.pl

Bartosz Rożek
Warsaw University of Technology
01142140@pw.edu.pl

Konrad Welkier
Warsaw University of Technology
01144707@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This proposal encapsulates the essential components of the 'Products Opinions and News' project, conducted as part of the NLP Course at MiNI PW Faculty. It covers crucial aspects such as defining the scientific goal, emphasizing the project's significance, detailing the concept and work plan, discussing the approach and research methodology, and justifying the project literature. It aims to elucidate the primary objectives, methodologies, and anticipated contributions of the project to the domain of Natural Language Processing.

1 Introduction

This project proposal covers the following information about the 'Products opinions and news' project implemented as part of the NLP Course at MiNI PW Faculty :

1. the scientific goal of the project (description of the problem to be solved, research questions and hypotheses);
2. significance of the project (state of the art, justification for tackling a specific scientific problem, justification for the pioneering nature of the project, the impact of the project results on the development of the research field and scientific discipline);
3. concept and work plan (general work plan - with a timeline, specific research goals, results of preliminary research, risk analysis);
4. approach & research methodology (underlying scientific methodology, methods, techniques and research tools, methods of results analysis, equipment and devices to be used in research);
5. well-justified project literature (a reference list for publications included in the project description, with full bibliographic data).

2 Literature review

Sentiment analysis, a fundamental aspect of natural language processing, has garnered significant attention due to its applications across various domains. Traditionally, sentiment analysis primarily focused on determining the overall sentiment polarity of text, classifying it as positive, negative, or neutral. However, the evolving landscape of sentiment analysis has shifted towards a more nuanced approach known as aspect-based sentiment analysis (ABSA).

Studies by (Liu and Zhang, 2012) laid the groundwork for ABSA, introducing the idea of exploring sentiments toward specific aspects or features within a text. This paradigm shift led to the exploration of sentiment analysis beyond document-level polarity, allowing for a more granular understanding of opinions within different aspects or entities mentioned in the text.

Recent studies by (Pontiki et al., 2014) have explored aspect-based sentiment analysis methodologies in English. They introduced techniques like aspect extraction and sentiment classification for fine-grained sentiment analysis, laying the groundwork for subsequent research in this field.

Sentiment analysis is a task highly connected to the language of the given text. This challenge can be addressed by creating multilingual models for such a task, such as a corpus proposed by (Augustyniak et al., 2023),

3 Goal of the project

The project's goal centers around evaluating sentiment analysis within product-related news articles, encompassing sentiments at both the comprehensive text level and their individual components. This includes discerning sentiment variations toward distinct product attributes and initially recognizing products and their mentioned features within the provided news articles.

Additionally, the project aims to create a Polish dataset and develop methodologies for extracting product names, their associated brands, models, and other pertinent parameters. The focus lies on delineating sentiment, particularly in user opinions and news texts, through an aspect-based sentiment analysis approach.

Moreover, the project endeavors to present sentiment results intelligently. It will involve a comprehensive exploration of the current State-of-the-Art (SOTA) techniques in sentiment analysis for both short and long texts. This exploration includes identifying aspects or features influencing sentiment, as well as delving into a spectrum of emotions or nuanced emotional scales. The overarching objective encompasses pinpointing sentiment-specific keywords and visualizing sentiment-related terms and aspects, potentially employing explainability techniques for enhanced clarity.

4 Significance of the project

The significance of this project is underscored by the existing void in the realm of Polish language datasets specifically tailored for aspect-based analysis in sentiment evaluation. Presently, there are available English datasets like Amazon Reviews (McAuley et al., 2015) and SemEval-2014 (Pontiki et al., 2014), which cater to aspect-based analysis, and one Polish dataset, PolEmo (Kocoń, Zaśko-Zielińska and Miłkowski, 2019), primarily designed for general sentiment analysis.

However, there is a noticeable absence of a dedicated Polish dataset geared towards aspect-based sentiment analysis. The creation of such a dataset constitutes a pivotal contribution, as it fills a crucial gap in the domain. This endeavor will pave the way for researchers, developers, and practitioners in the field to explore and advance their methodologies in aspect-based sentiment analysis within the Polish language.

The initiation of this dataset serves as a fundamental enabler, allowing for the development and testing of innovative approaches in the realm of aspect-based sentiment analysis specifically tailored to the nuances and intricacies of the Polish language. Its creation is poised to invigorate research and application within the domain, opening avenues for diverse applications and further enhancing the understanding of sentiment analysis within the Polish linguistic context.

5 Concept and work plan

5.1 Concept

The output value of the project would be models for three different areas:

- Sentiment analysis in the article/reviews domain
- Aspect-based sentiment analysis in the article/reviews domain
- Extraction of the products' information from the articles and reviews

All of these approaches are more deeply described in the 6 section.

5.2 Work plan

The work plan is divided into phases:

1. **Preparation** - literature and state-of-the-art analysis - this phase is already done and a deliverable of it is this report.
2. **Data preprocessing** - obtaining, cleaning and adjusting data for the sentiment analysis exercise
3. **Development preparation** - the process of creating a plan for the framework which will allow us to incorporate ready tools for our purposes in one structure
4. **Development** - strictly programming phase where we will implement tools using state-of-the-art solutions
5. **Testing** - comparing results of different approaches

6. **Creating Polish data set** - using prepared models we will transfrom the PolEmo dataset to extract aspects from the reviews and create an aspect-version from it.

5.3 Timeline

The timeline of the project is strictly connected to the timeline of the classes of NLP course. Counting 25.10 - 1.11 as a week 1, the timeline is as follows:

1. **Preparation** - week 1 & week 2 (25.10-5.11)
2. **Data preprocessing** - week 3 (6.11-12.11)
3. **Development preparation** - week 3 (6.11-12.11)
4. **Development** - week 3 & week 4 (6.11-19.11)
5. **Testing** - week 5 (20-26.11)
6. **Creating Polish data set** - week 6 (27.11-3.12)

5.4 Research goals

The main goals of the project are preparing a set of different approaches to one problem and testing them. We will incorporate many different tools that will collaborate to create additional value.

5.5 Preliminary research

Preliminary research showed two main problems in the project:

- **Data** - there are not many annotated data sets for aspect-based sentiment analysis, and we did not manage to find one in Polish. Because of these facts, we decided to overcome that problem with the usage of the translator on the Polish data set.

- **End-to-end solutions** - the state-of-the-art world does not provide us with many high-performance tools that can be easily used in aspect-based sentiment analysis. Thus, we decided to merge the functionalities of many different tools.

5.6 Risk analysis

In this project, we can encounter some problems, such as:

- Lack of time - we want to stick to the timeline and even try to speed up a bit.

- Insufficient memory - to overcome this problem we decided to use sampled data and treat this project as a ”Proof of concept” that may be developed in the future.

- ChatGPT API limitations - the free tier access is limited in the number of requests

6 Approach and research methodology

6.1 Datasets

- **SemEval 2014 Task 4** - The SemEval 2014 Task 4 dataset (Pontiki et al., 2014) is a benchmark dataset used to evaluate systems for aspect-based sentiment analysis. The task is divided into four subtasks:
  - **Subtask 1: Aspect Term Extraction** - Extract the explicit aspect term (e.g., ”battery life”) from the sentence.
  - **Subtask 2: Aspect Term Polarity** - Determine the sentiment polarity of the aspect term mentioned in a sentence.
  - **Subtask 3: Aspect Category Detection** - Identify the category of the aspect (e.g., ”food”, ”service”) that is mentioned in a sentence
  - **Subtask 4: Aspect Category Polarity** - Determine the sentiment polarity for the aspect category.

- **Amazon Reviews (Electronics Subset)** - The Amazon Reviews dataset (McAuley et al., 2015) is a collection of reviews written by customers for products purchased on Amazon. It’s one of the largest and most commonly used datasets for sentiment analysis and natural language processing tasks. This dataset includes reviews spanning various product categories, providing a broad range of vocabulary and topics. We will be using the *Electronics* subset of the Amazon reviews dataset as it is a very large dataset, and the authors themselves encourage to make use of a subset.

- **PolEmo** - The PolEmo dataset This dataset was instroduced in a paper (Kocoń, Zaśko-Zielińska and Miłkowski, 2019). It consists of Polish reviews from four domains: medicine, hotels, products, and school. Each review is labelled with one of the following labels: ”negatywny”, ”ambiwalnetny”, ”neutralny”, and ”pozytywny” (negative, ambivalent, neutral, and positive)

6.2 Tools

- **ChatGPT** - ChatGPT (2021) is a state-of-the-art tool, that is perfect for processing text data, and it can be leveraged to do multiple things. We will use it to perform aspect-based sentiment annotations, divide sentences into chunks to use with Flair, and extract keywords from sentences to use with Sentistrength

- **Flair** - Flair (2019) is a simple-to-use framework for performing NLP Tasks such as Named Entity Recognition and Sentiment Analysis.

- **SpaCy** - SpaCy (2017) is a free, open-source library for advanced Natural Language Processing (NLP) in Python. It can be used to perform Tokenization, Part of speech tagging, NER and other NLP tasks.

- **Sentistrength** - Sentistrength (2010) is a text analysis tool designed specifically for sentiment analysis. It is capable of assigning a sentiment strength score to text, which is useful for detecting positive and negative sentiment in short texts, even for informal language often found in product reviews.

- **BERT** - BERT (2018) (Bidirectional Encoder Representations from Transformers) is a groundbreaking method in Natural Language Processing (NLP) that was introduced by researchers at Google AI in 2018

- **pyABSA** - pyABSA (2023) is a open source Framework for Aspect-based Sentiment Analysis. It allows both - extracting aspects as well as labeling them with sentiment.

6.3 Approaches

As mentioned in the previous section, the project is divided into three ”subtasks”, all of them are presented below.

6.3.1 Sentiment analysis

- **ChatGPT** - the whole process of analysis will be handled by Chat. A suitable prompt will be tested to get a response that will match expectations.

- **Flair** - Flair has an out-of-the-box feature of assigning sentiment labels with a probability of them being true.

6.3.2 Aspect-based sentiment analysis

- **Chat GPT** - just as in the previous approach, the well-constructed prompt to Chat GPT will allow us to get results in the appropriate form.

- **Flair/Spacy + SentiStrength** - analysis will be divided into two parts
  - extract keyword using Flair/Spacy
  - analyze text with given keywords with SentiStrenght

- **ChatGPT + SentiStrength** - similarly to the previous approach but Chat GPT will do the extraction of the keywords

- **ChatGPT + Flair** - again, the task is divided into two parts
  - divide the text into chunks using Chat GPT
  - analyze the sentiment of each chunk using Flair

- **BERT** - Bert will perform the whole process of analyzing

- **pyABSA** - pyABSA is capable of performing the whole flow of aspect-based analysis - from extracting aspects to assigning the sentiment to them

6.3.3 Extraction of information

This task is the most inconvenient one - there are no suitable data sets for this kind of problem. This will be handled with the use of Chat GPT and checking whether the results seem correct.

7 Team Members Contribution

Anish Gupta	Research, a part of the report, report checkup
Martyna Majcharzak	Report abstract, introduction, literature review, goal and significance of the project
Bartosz Rożek	Research, a part of the report preparation, presentation preparation & delivery
Konrad Welkier	datasets preparation and Exploratory Data Analysis, presentation preparation & delivery report check-up

8 Rebuttal

The main concerns from the reviews was the lack of a formal literature review and lack of clarity about the way the Polish Dataset was supposed to be created.

We have included the literature review in this revised version of the project proposal and fixed some of the pointed-out problems with the references.

The problem with the Polish Dataset was caused by problems in communication with the external firm, that was supposed to deliver the data. Due to that we decided to abandon the idea of using this dataset and use the PolEmo dataset instead. We have included the description of this dataset in this version of the report.

References

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy and Tomasz Kajdanowicz. 2023. *Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark.*

Bing Liu and Lei Zhang 2012. *A SURVEY OF OPINION MINING AND SENTIMENT ANALYSIS.*

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. *Image-based recommendations on styles and substitutes.* In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 43–52.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. *SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Kocoń, Jan and Zaśko-Zielińska, Monika and Miłkowski, Piotr 2019. *PolEmo 2.0 Sentiment Analysis Dataset for CoNLL*, CLARIN-PL digital repository

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv preprint arXiv:1810.04805.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. *FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP.* In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. *Sentiment Strength Detection in Short Informal Text.* Journal of the American Society for Information Science and Technology, 61(12):2544–2558.

OpenAI. 2021. *ChatGPT: Optimizing Language Models for Dialogue.* `https://openai.com/blog/chatgpt`.

Matthew Honnibal and Ines Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* To appear.

Matthew Honnibal and Ines Montani. 2023. *PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis.*

@miscyang2023pyabsa, title=PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis, author=Heng Yang and Chen Zhang and Ke Li, year=2023, eprint=2208.01368, archivePrefix=arXiv, primaryClass=cs.CL