UDC 159.942:004.42

7. SENTIMENT ANALYSIS OF LINUX KERNEL SUBSYSTEMS

Ramanouski M.D., Bachelor Degree Student, gr.353501

*Belarusian State University of Informatics and Radioelectonics
Minsk, Republic of Belarus*

Ladyjenko M.V. – Senior Lecturer

Annotation. The article has a purpose to present the concept of sentiment analysis. The classification of sentiments (positive, negative, or neutral) is presented. The paper considers the differences between communication styles of developers of different Linux kernel subsystems. It explores the use and correlation between automated sentiment and emotion detection tools. Emotional scoring is employed to evaluate written language content.

Keywords. Sentiment analysis, emotional scoring, open-source software development, Linux kernel, automated emotion analysis.

Sentiment analysis is a field of study that uses computational methods to analyse, process, and reveal people's feelings, sentiments, and emotions hidden behind a text or interaction. It uses machine learning (ML), natural language processing (NLP), data mining, and artificial intelligence (AI) techniques to mine, extract and categorise users' opinions on a company, product, person, service, event, or idea for various sentiments. Sentiment analysis also referred to as opinion or sentiment mining, captures the polarity of the text, which often falls under the categories of positive, negative, or neutral. Moreover, associating sentiments and emotions with text runs across different levels, such as sentences, paragraphs, and documents [1]. As modern development teams are global, they communicate over digital platforms such as EMail, Slack and Gitter. These communication channels are an ideal source for automated opinion mining to provide feedback on feelings of developers and improve productivity of teams.

Open source software development highly depends on the effectiveness of communication between developers. As a consequence, positivity in communication channels can improve the development process, while negativity is highly likely to scare off new contributors [2]. Large open source projects are often divided into smaller specialised subsystems with different participants. As a result, there might be differences in the styles of communication of subsystems inside a large project. The goal of this paper is to analyse human language emotional connections and nuances by using automated sentiment analysis tools and to assign a numerical score or category to the emotional content, allowing for a more objective and standardised assessment. The Linux kernel was chosen to test this theory as it is a big project with a wide variety of subsystems and contributors.

The Linux kernel employs a number of mailing lists to facilitate communications among developers [3]. Seven mailing lists covering different subsystems selected for this study are displayed in Table 1.

Table 1 – Subsystems selected for the study

Mailing List	Mailing list description	Selection criteria
lkml	General list that includes most messages	Provides a good average across whole project
dri-devel	Graphical Processing Unit's (GPU) subsystem	Large critical subsystems with high number of participants
netdev	Networking subsystem	
linux-fsdevel	File system subsystem	
kvm	Virtualisation subsystem	
intel-gfx	Intel GPU driver development	Mainly used by a singular company that might impose its own communication standards
rust-for-linux	Integration of Rust into the Linux kernel	Discussion of integrating second innovative programming language, bringing in developers from completely different fields

The archives of the mailing lists selected in Table 1 were downloaded from the official sources. The emails presented in these lists include additional information, such as code patches. To improve the quality of analysis, messages were additionally filtered and preprocessed to leave out only verbal means of communication. First, non-reply messages and messages generated by automatic tools were filtered out.

Secondly, the messages were preprocessed by removing quotes of other emails, code patches, links and email addresses, using similar approaches as described in other papers that analyse Linux mailing lists [4]. This process resulted in a clean set of 5,000 emails for each subsystem, with the exception of rust-for-linux list where due to its novelty only 3,250 messages were collected. The preprocessed mails were analysed by using Senti4SD, a sentiment polarity classifier for software developers' artifacts [5]. Senti4SD rates messages as negative, neutral or positive. It is built by employing a distributional semantic model which is trained on posts and comments from StackOverflow. This helps to improve the quality of classification for software-related texts compared to general tools, such as SentiStrength and NLTK Vader. The results of running Senti4SD are presented in Table 2. To provide better understanding the positive and negative data is visualised in Figure 1.

Table 2 – Results of sentiment analysis with Senti4SD

Subsystem	Negative	Neutral	Positive
lkml	21.68 %	39.68 %	38.64 %
dri-devel	19.72 %	43.32 %	36.96 %
netdev	18.92 %	34.74 %	46.34 %
linux-fsdevel	28.12 %	40.70 %	31.18 %
kvm	26.88 %	39.48 %	33.64 %
intel-gfx	22.10 %	44.80 %	33.10 %
rust-for-linux	23.12 %	41.99 %	34.89 %

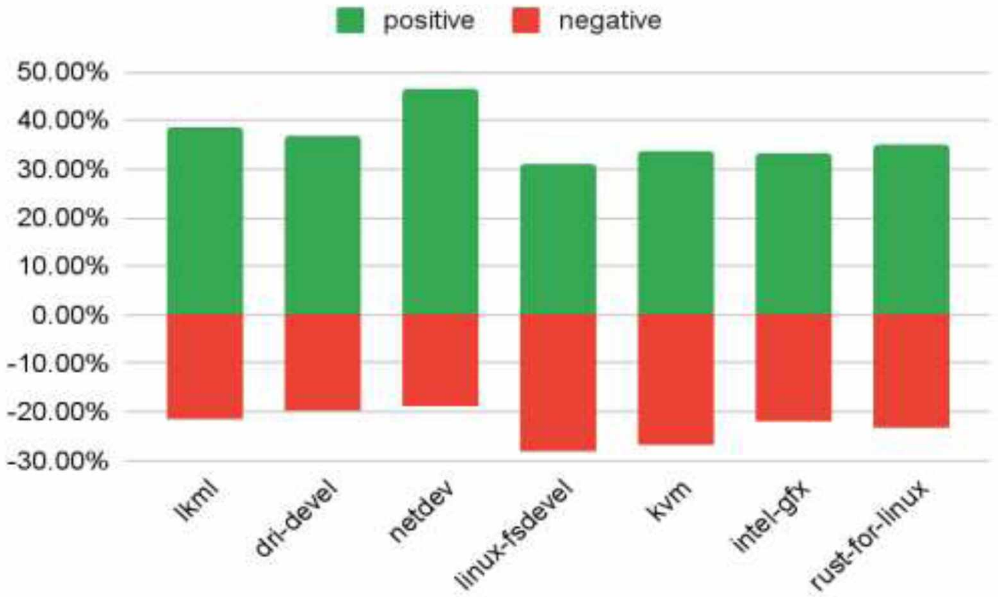

Figure 1 – Positive and negative messages plotted in a bar chart

As it can be seen in Figure 1, while subsystems share similar scores with a difference of ±5 %, there are some outliers. Netdev mailing list contains more positive messages and less negative, while kvm and fsdevel contain generally more negative messages. However, these differences are mild and should not drastically affect new contributors to choose one subsystem over another.

Another interesting area to consider is emotions expressed through developers' emails. To analyse that, the RoBERTa, a deep learning transformer model, fine-tuned on emotions dataset, is used [6]. The model receives a text and produces multiple pairs of emotions and scorings, i.e. how sure the model is that the emotion is present. To acquire the overall emotional scoring of the mailing list, top 3 emotions from each message are extracted using this model, and then the average for each emotion is calculated. This process provides emotional scoring for each mailing list. After performing this analysis, the six most prevalent emotions were identified: neutral (no emotion present), gratitude, confusion, approval, curiosity and disapproval. Together these emotions make up more than 85 % of mailing lists emotions. The results of this emotional scoring are presented in Figure 2.

As it can be seen in Figure 2, the most prevalent emotional scoring among subsystems are quite similar. The netdev mailing list has the highest gratitude percentage, while linux-fsdevel and kvm have the highest neutral rates. A strong correlation between "gratitude" detected by RoBERTa and "positivity" identified by Senti4SD with a coefficient of 0.8139 was found.

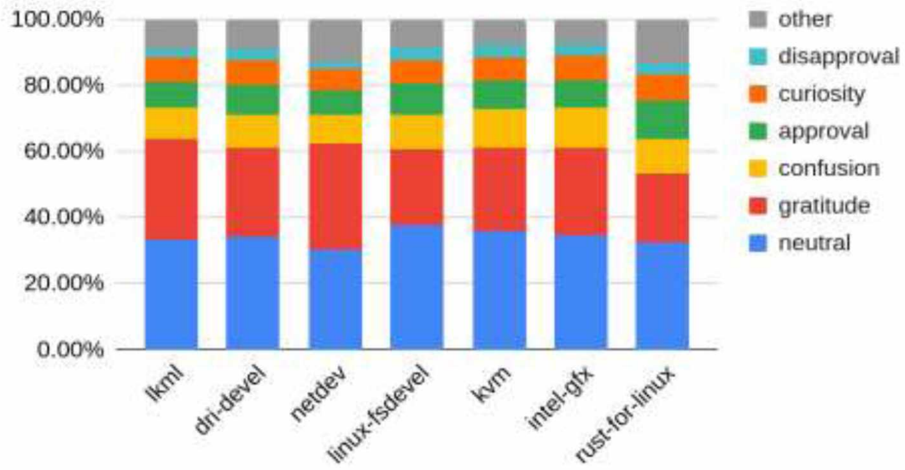

Figure 2 – Most prevalent emotions

The less prevalent emotions that have share of about 1–2 % are presented in Figure 3. According to the results in Figure 3, there is an outlier of rust-for-linux which has surprisingly higher levels of "optimism" and "joy", compared to other mailing lists. This reflects the overall optimistic view of rust-for-linux by developers of kernel [7].

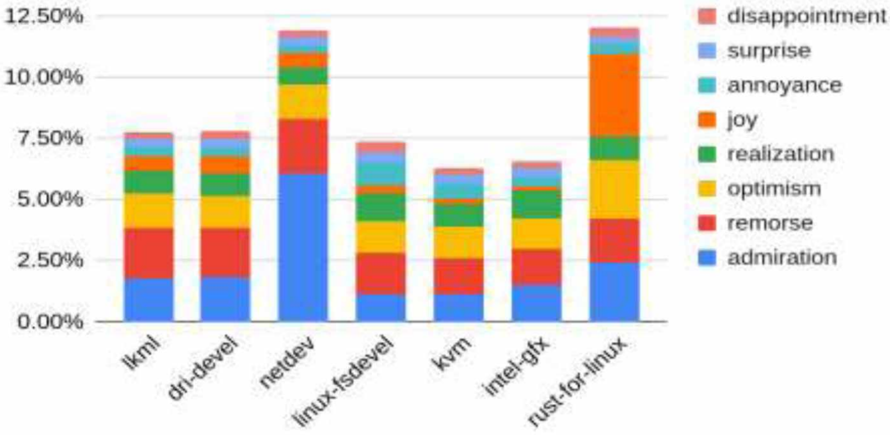

Figure 3 – Less prevalent emotions

The findings of this paper suggest that there are minor differences between different Linux kernel subsystems in terms of sentiment (positive, negative, or neutral) behind a piece of communication between developers. It is important to state that netdev subsystems are usually the most positive ones while kvm and linux-fsdevel are the least. Furthermore, another finding of this paper shows the overall positive and optimistic view on the new rust subsystem in kernel. Finally, the provided evidence indicates that there is a correlation between general emotion analysis tools such as RoBERTa and software-specific sentiment classification tools such as Senti4SD.

References:

1. What is Sentiment Analysis? [Electronic resource]. – Mode of access: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-sentiment-analysis/. – Date of access: 14.03.2024.

2. Code of Conduct in Open Source Projects [Electronic resource]. – Mode of access: https://www.researchgate.net/publication/. – Date of access: 11.03.2024.

3. Love, R. Linux Kernel Development, second edition / Robert Love. – Boston ; *London : Pearson Education, 2010. – 396 p.*

4. A Longitudinal Study on the Maintainers' Sentiment of a Large Scale Open Source Ecosystem [Electronic resource]. – Mode of access: https://www.researchgate.net/publication/335644688. – Date of access: 12.03.2024.

5. Sentiment Polarity Detection for Software Development [Electronic resource]. – Mode of access: https://dl.acm.org/doi/10.1145/3180155.3182519/. – Date of access: 09.03.2024.

6. GoEmotions: A Dataset of Fine-Grained Emotions [Electronic resource]. – Mode of access: https://www.researchgate.net/publication/343300732/. – Date of access: 11.03.2024.

7. Rust in the Linux Kernel [Electronic Resource]. – Mode of access: *https://thenewstack.io/rust-in-the-linux-kernel/.* – Date of access: 14*.03.2024.*