**MINOR** ​ **PROJECT**​​ **REPORT**​

# SENTIMENT​ ​ANALYSIS

Submitted ​ in​ ​ partial​ ​ fulfilment​ ​ of​ ​ the​ ​ requirements​ ​ for​ ​the​ ​award​ of​ ​the​ ​degree​ ​of​ Bachelor​ of​ ​ Technology​ ​ in

**Information**​​**Technology**

**Under**​​**the**​​**guidance**​​**of**

Prof.​ ​Arvinder​ ​Kaur

&

Mr.​ ​Amrit​ ​Pal​ ​Singh

**Submitted** ​ **by**

Divesh Bisht

(04616401514)

7​th​​ ​Semester



**University**​​**School**​​**of**​​**Information**​​**and**​​**Communication**​​**Technology**

**GGS**​​**Indraprastha**​​**University,**​​**Delhi-78**

**CERTIFICATE**

This is to certify that the project titled, “Sentiment Analysis" submitted by "Divesh Bisht" in partial fulfilment of the requirements for the award of "Bachelor of Technology" in "Information Technology" at the "University School of Information and Technology, GGS Indraprastha University, Delhi-78" is an authentic work carried out by him under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree or Diploma.

Date:

Mrs Arvinder Kaur   
(Dean, USICT)

# ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to our college University

School of Information and Communication Technology, GGS Indraprastha University, Dwarka, Delhi for providing me an opportunity to undertake and complete such an interesting project and also to everyone who supported me throughout the course of this project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number​ ​of​ ​issues​ ​related​ ​to​ ​the​ ​project.

I am very grateful to our respected faculties and my project mentors**,**​ **Mrs.**

**Arvinder Kaur (Professor & Dean USICT)** ​and ​**Mr. Amrit Pal Singh (Assistant Professor)** ​for their continuous support and guidance throughout​this project.

I am also very thankful to my partner, Guneet who had helped me a lot in completing​ ​this​ ​project.

Thank​ ​you,

Divesh​ ​Bisht.

**Contents**

[SENTIMENT​ ​ANALYSIS 1](#_Toc498331891)

CERTIFICATE……………………………………………………………………. 2

[ACKNOWLEDGEMENT 3](#_Toc498331892)

[ABSTRACT 5](#_Toc498331893)

[1. ​ INTRODUCTION​ 6](#_Toc498331894)

[2. ​ RELATED​ ​ WORK​ 7](#_Toc498331895)

[3. ​ RESEARCH​ ​ METHODOLOGY​ 8](#_Toc498331896)

[3.1 ​ DATASET​ 8](#_Toc498331897)

[3.2 ​ EMOTION​ ​ MINING​ 9](#_Toc498331898)

[3.3 ​ SENTIMENT​ ​ ANALYSIS​ ​ TOOLS​ 10](#_Toc498331899)

[3.4 ​ PERFORMANCE​ ​ MEASURES​ 11](#_Toc498331900)

[4. RESULTS​ 13](#_Toc498331901)

[5. CONCLUSIONS​ 15](#_Toc498331902)

[6. REFERENCES​ 17](#_Toc498331903)

**ABSTRACT**

Sentiment analysis refers to the use of natural​ language processing, ​ text​ analysis​, computational​ linguistics,​ and biometrics​ to​ systematically identify, extract, quantify, and study affective states and subjective information. ​It is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. Recent years have witnessed an increasing attention to social aspects of software engineering. This includes studies of emotions and sentiments which are experienced and expressed by the software developers.

We present a computer-assisted literature review where we utilize both text mining and qualitative coding to answer whether developers feel emotions. The study is carried out on our dataset which consists of comments posted by software developers​ ​on​ ​software​ ​version​ ​maintainence​ ​website​ ​Apache​ ​Jira.

We study whether the sentiment analysis tools agree with the sentiment recognized by human evaluators. We also study to what extent the results obtained from different sentiment analysis tools agree with each other. Furthermore, we evaluate the impact of the choice of a sentiment analysis tool on software engineering studies​ ​for​ ​positive,​ ​negative​ ​and​ ​neutral​ ​texts.

# ​ INTRODUCTION​

The study consists of two prime processes that is Emotion Mining and Sentiment Analysis. Emotion mining tries to identify the presence of human emotions like joy, love, sadness, fear etc from text, voice and video artifacts produced by humans As such, it is different from sentiment analysis, which instead evaluates a given emotion as being positive or negative[9]. Absence of emotion is marked as neutral sentiment.

Recently, software development field has become increasingly social. Software Developers interact with each other through means of issue reports, posting updates on some version maintainance website or software or blogging about their work and progress. These platforms have numerous amount of comments in which developers have shown and expressed their emotions towards other developers and towards components of an in progress software. Their emotions play an important role during the life cycle of a software and proper analysis of their sentiments and emotions might help them in enhancing the efficiency of the SDLC and hence the software itself. Therefore, it has become necessary to know if software developers feel emotions i.e. whether emotions can be extracted from comments posted by developers​ on​​ ​software​ ​development​ ​ecosystems.

Our​ ​study​ ​was​ ​divided​ ​into​ ​5​ ​stages​ ​which​ ​are ​​described​ ​below​ ​:

* Finding and understanding previously published research papers that deal with​ ​the​ ​issue​ ​of​ ​emotion​ ​mining​ ​and​ ​sentiment​ ​analysis.
* Collecting data through python scripts and cleaning the data obtained so that it ​ can​ ​ be​ ​ processed​ ​ easily.​
* Manually mining emotions from the collected data and grouping various emotions​ ​into​ ​positive,​ ​negative​ ​or​ ​neutral​ ​sentiments.
* Sentiment analysis of the dataset carried out using Natural Language Processing tools such as NLTK, Sentistrength and WatsonNLU( Natural Language​ ​Understanding​ ​).
* Calculating metrics weighted kappa and adjusted rand index for evaluating agreement​ ​between​ ​manual​ ​labelling​ ​and​ ​sentiment​ ​analysis​ ​tools.

# ​ RELATED​ ​ WORK​

Murgia et al. analysed the emotions present in software developer comments by labelling​ ​them​ ​manually​ ​through​ ​evaluators[1].​ ​There​ ​was​ ​no​ ​use​ ​of​ ​tools​ ​involved.

Jongeling et al. further performed sentiment analysis on the dataset made available by Murgia’s study and calculated various agreement statistics on the results obtained[2].

We​ ​try​ ​to​ ​address​ ​the​ ​following​ ​research​ ​questions​ ​through​ ​our​ ​study:

*RQ1. Do human raters agree on the presence or absence of emotions in issue reports?*

*RQ 2. To what extent do different sentiment analysis tools agree with emotions of software*​​*developers?*

*RQ 3. To what extent do results from different sentiment analysis tools agree with each*​​*other?*

*RQ 4. To what extent the intersection of different sentiment analysis tools agree with* ​​*emotions*​​*of*​​*software*​​*developers?*

# ​ RESEARCH​ ​ METHODOLOGY​

This section discusses the dataset used in our analysis, the general procedure used to rate extracted comments and description of various sentiment analysis tools and agreement​ ​statistics​ ​used.

## ​ DATASET​

Having hands on experience with Python, we automated the whole process. Python scripts were written using *BeautifulSoup*​ and *Urllib*​​libraries in order to extract the software developers’ comments along with username, time of comment and issue ID from ​**Apache Jira’s** open source projects. For the sake of simplicity, we took only three open source projects in consideration i.e. MTOMCAT, RAT and LUCENE.​ ​We​ ​extracted​ ​1117​ ​comments.

A total of 500 comments were included in the final dataset and all the code snippets in the comments were removed. These comments were taken in the groups of ten or more so as to maintain the context flow and help human raters familiarize themselves with the ongoing conversations so as to get an abstruse information about​ ​the​ ​emotions​ ​present.

## ​ EMOTION​ ​ MINING​

For the whole emotion mining process, the complete dataset of 500 comments was distributed among 4 evaluators. Each evaluator rated the particular issue comment as​ ​having​ ​a​ ​particular​ ​emotion​ ​or​ ​nil​ ​in​ ​case​ ​no​ ​emotion​ ​was​ ​found.

We used Parrott’s six primary emotions i.e. love, joy, surprise, anger, sadness and fear. The raters were given a guide containing 2-3 examples of each emotion to help​ ​them​ ​understand​ ​the​ ​labeling​ ​process[1].

Emotion:​ ​Love

* Thanks​ ​for​ ​your​ ​input!​ ​You’re​ ​like​ ​awesome.
* Thanks​ ​very ​ much!​ ​ I​ ​ appreciate​ ​ your​ ​ efforts.​
* Would​ ​love​ ​any​ ​advice.

Emotion:​ ​Joy

* Great​ ​work ​ you​ ​ guys!​
* Hope​ ​this​ ​will​ ​help​ in​ ​​identifying​ ​more ​ usecases.​
* I’m​ ​happy​ ​with​ ​the​ ​approach​ and​ ​​the​ ​code​ ​looks​ ​good.

Emotion:​ ​Surprise

* I​ ​still​ ​question​ ​the​ ​default,​ ​which​ ​can ​​lead ​ to​ ​ surprisingly​ ​ huge​​ ​memory​ loss.​
* Oops.​ ​It​ ​needs​ ​to​ ​be​ ​added​ ​to​ ​Makefile.
* I​ ​also​ ​documented​ ​an​ ​unexpected ​feature​​ ​with​ ​the​ ​ServletResolver.

Emotion:​ ​Anger

* I​ ​will​ ​come​ ​over​ ​to​ your​ ​​work​ and​ ​ slap​ ​ ​you.

● This​ ​is​ ​an​ ​ugly​ ​workaround.

* WTF,​ ​a​ ​package​ ​refactoring​ ​and​ class​ ​​renaming​ ​in​ ​a​ ​patch.

Emotion:​ ​Sadness

* Sorry​ ​for​ ​the​ ​delay.
* Wish​ ​I​ ​had​ ​payed ​ more​​ attention​​ ​in​ ​my​ ​Java​ ​class.
* No​ ​need​ ​to​ ​fix​ ​this,​ as​ ​​sad​ ​as​ ​it​ ​is.

Emotion:​ ​Fear

* I’m​ ​most​ ​concerned​ ​with​ some​ ​​of​ the​ ​ timeouts.​
* I​ ​suspect​ ​that​ ​remove​ ​won’t ​work​​ ​in ​ this​ ​ case.​
* I’m​ ​worried​ ​about​ ​some​ ​subtle​ differences​ ​​between​ ​char​ ​and​ ​Character.

Once all the labeling was done, final mapping values which indicated whether a given comment is positive, negative or neutral were obtained using the following guidelines[2]:

* We consider the comment as positive if at least three evaluators have indicated a positive sentiment and no evaluator has indicated negative sentiments.
* We consider the comment as negative if at least three evaluators have indicated a negative sentiment and no evaluator has indicated positive sentiments.
* We consider the comment as neutral when three or more evaluators have neither​ ​indicated​ ​a​ ​positive ​ sentiment​ ​ nor​ ​ a​ ​ negative​ ​ sentiment.​

● Comment​ ​is​ ​considered​ ​contradictory​ ​for​ ​all​ ​remaining​ ​cases.

There were 2 comments for which 3 or more evaluators had given ‘surprise’ as emotion. These were removed from the dataset since surprise can be treated as having​ ​both​ ​positive​ ​and​ ​negative​ ​sentiments[2].

We found that 112 comments have been labeled contradictory and they were also removed​ ​from​ ​the​ ​dataset​ ​leaving​ ​500-2-112​ ​=​ ​386​ ​comments​ ​in​ ​the​ ​dataset.

## ​ SENTIMENT​ ​ ANALYSIS​ ​ TOOLS​

We did not employ tools that required training before they could be applied. We chose​ ​the​ ​following​ ​three​ ​tools​ ​to​ ​perform​ ​sentiment​ ​polarity​ ​analysis:

* **NLTK** - NLTK has been used in previous sentiment analysis studies[2],[5]. It uses a simple bag of words model and returns for each text three probabilities: a probability of the text being negative, one of it being neutral and one of it being positive. If neutral is greater than 0.5 then the label will be neutral. Otherwise, the label will be positive or negative, whichever has the greater probability. We use the API provided at text-processing.com to use​ ​NLTK.

* **SENTISTRENGTH** ​– SentiStrength is a sentiment analysis tool and generally first choice of researchers[2]. It assigns an integer value between 1 and 5 for the positivity of a text and a value between −1 and−5 for the negativity, n. To obtain mapping values from these values, we followed Thelwall et al[4]. A text is considered positive when (p+n)>0 , negative when​ ​(p+n)<0​ ​and​ ​neutral​ ​if​ ​p+n=0​ ​&​ ​p>=4.

* **WATSON NLU –** The​ API returns for a text fragment a score which is in the range [−1,1]. For negative scores, the type is negative, positive for positive scores and neutral for 0. The status reflects the analysis success and it​ ​is​ ​either​ ​“OK”​ ​or​ ​“ERROR”.

## ​ PERFORMANCE​ ​ MEASURES​

Out of 386 comments, 11 showed ERROR status with the WatsonNLU and hence they were removed from the dataset. Confusion matrices were calculated on remaining​ ​375​ ​comments[2].

Since more than 58% comments were neutral, so traditional metrics such as accuracy might be misleading as the dataset is unbalanced. Thus, we go for weighted​ ​kappa​ ​and​ ​ARI[3],[8].

**Weighted kappa** (κ) as recommended by Bakeman and Gottman is a measure of interrater​ ​agreement[3].​ ​The​ ​agreement​ ​due​ ​to​ ​chance​ ​is ​ factored​ ​ ​out​ ​using​ ​kappa.

Since the sentiments can be seen as ordered, from positive through neutral to negative, and disagreement between positive and negative is more severe than between positive and neutral or negative and neutral, so the weighting scheme we followed​ ​is​ ​as​ ​shown​ ​in​ ​Table[3].

Weighting​ ​Scheme​ ​for​ ​Weighted​ ​Kappa​ ​Computation

|  |  |  |  |
| --- | --- | --- | --- |
|  | Neutral | Positive | Negative |
| Neutral | 0 | 1 | 1 |
| Positive | 1 | 0 | 2 |
| Negative | 1 | 2 | 0 |

We​ ​considered​ ​the​ ​explanation​ ​of​ ​Kappa​ ​as​ ​suggested​ ​by​ ​Viera​ ​and​ ​Garrett[7].

Interpretation​ ​of​ ​Kappa​ ​values

|  |  |
| --- | --- |
| **Kappa**​​**Value** | **Interpretation**​​**of**​​**Agreement**​​**between**​​**Raters** |
| <0 | Poor |
| 0​ ​-​ ​0.20 | Slight |
| 0.21​ ​-​ ​0.40 | Fair |
| 0.41​ ​-​ ​0.60 | Moderate |
| 0.61​ ​-​ ​0.80 | Substantial |
| 0.81 - 1.0 | Almost perfect |

​​​​ ​​

**Adjusted Rand Index** (ARI) by Hubert and Arabie corrects for the possibility that pairs of observations have been put in the same category by chance[8]. ARI evaluates​ ​whether ​ pairs​ ​ of​

comments are considered as belonging to the same sentiment rather than on whether​ ​the​ ​comments​ ​have​ ​been​ ​assigned​ ​to​ ​correct​ ​sentiment.

The expected value of ARI for independent partitions is 0 and 1 for identical partitions, the closer the value of ARI to 1 the better the correspondence between the​ ​partitions.

# RESULTS​

Confusion​ ​matrices​ ​for​ ​weighted​ ​kappa​ ​and​ ​ARI​ ​computation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | neu | pos | neg |  |  | neu | pos | neg |
| **NLTK** | **Manual** | |  |  | **SS** | **Manua**​l | |  |
| neu | 73 | 23 | 5 |  | neu | 134 | 35 | 17 |
| pos | 47 | 57 | 3 |  | pos | 35 | 58 | 3 |
| neg | 100 | 22 | 45 |  | neg | 51 | 9 | 33 |
|  |  | |  |  |  |  | |  |
| **WNLU** | **Manual** | |  |  | **NLTK** | **SS** | |  |
| neu | 102 | 18 | 6 |  | neu | 71 | 10 | 20 |
| pos | 25 | 71 | 3 |  | pos | 37 | 57 | 13 |
| neg | 93 | 13 | 44 |  | neg | 78 | 29 | 60 |
|  |  |  |  |  |  |  |  |  |
| **NLTK** | **WNLU** |  |  |  | **SS** | **WNLU** |  |  |
| neu | 66 | 12 | 23 |  | neu | 99 | 25 | 62 |
| pos | 26 | 65 | 16 |  | pos | 17 | 65 | 14 |
| neg | 34 | 22 | 111 |  | neg | 10 | 9 | 74 |
|  |  |  |  |  |  |  |  |  |
| **Manual** | **NLTK,SS** | |  |  | **Manual** | **NLTK,WNLU** | |  |
| neu | 57 | 11 | 30 |  | neu | 54 | 9 | 62 |
| pos | 12 | 46 | 3 |  | pos | 11 | 54 | 9 |
| neg | 2 | 0 | 27 |  | neg | 1 | 2 | 40 |
|  |  |  |  |  |  |  |  |  |
| **Manual** | **SS,WNLU** | |  |  | **Manual** | **NLTK,SS,WNLU** | | |
| neu | 80 | 11 | 42 |  | neu | 46 | 2 | 28 |
| pos | 16 | 53 | 4 |  | pos | 10 | 45 | 2 |
| neg | 3 | 1 | 28 |  | neg | 1 | 0 | 25 |

Agreement​ ​of​ ​sentiment​ ​analysis​ ​tools​ ​with​ ​manual​ ​labelling​ ​and​ ​with​ ​each​ ​other

|  |  |  |
| --- | --- | --- |
| **TOOLS** | **KAPPA** | **ARI** |
| NLTK​ ​vs​ ​Manual | 0.3 | 0.06 |
| SentiStrength​ ​vs​ ​Manual | 0.39 | 0.139 |
| WNLU​ ​vs​ ​Manual | 0.43 | 0.177 |
| NLTK​ ​vs​ ​SentiStrength | 0.3 | 0.069 |
| NLTK​ ​vs​ ​WNLU | 0.49 | 0.221 |
| SentiStrength vs WNLU | 0.49 | 0.202 |

Agreement​ ​of​ ​groups​ ​of​ ​sentiment​ ​analysis​ ​tools​ ​with​ ​manual​ ​labelling

|  |  |  |
| --- | --- | --- |
| **TOOLS** | **KAPPA** | **ARI** |
| NLTK,​ ​SentiStrength | 0.6 | 0.297 |
| NLTK,​ ​WNLU | 0.51 | 0.24 |
| SentiStrength,​ ​WNLU | 0.55 | 0.283 |
| NLTK,​ ​SentiStrength,​ ​WNLU | 0.66 | 0.405 |

# CONCLUSIONS​

**RQ1:** ​When identifying emotions, out of 500 comments, only in **29.2**​ **%** comments all the four raters had the same rating. From these 146 comments, 117 showed absence of emotion and were marked nil by all the four raters. From the rest 29, 17 were of joy, 7 of sadness, 2 of surprise and 1 each of love, anger and fear. When identifying sentiments, the two comments marked ‘surprise’ by all the four evaluators were removed as surprise can be treated​ ​as​ ​both​ ​negative​ ​and​ ​positive.

Out of remaining 498 comments, only in **42.9​ %** comments all the four raters had identified same emotion i.e. positive, negative or neutral. From these 214 comments, 123 were marked as neutral, 15 were marked negative (marked either sadness, anger or fear) and rest 76 were marked positive (marked either​ ​love​ ​or​ ​joy).

To conclude, we can say that raters agreed on absence of emotions more as compared​ ​to​ ​presence​ ​of​ ​emotions.

**RQ2:** Results clearly indicate that the sentiment analysis tools do not agree with the manual labelling as no tool can achieve substantial or better agreement value of kappa (0.6 or more). Only moderate and fair agreement was obtained between tools and manual labelling. WatsonNLU scores best with kappa=0.43 and

ARI=0.177,​ ​followed​ ​by​ ​SentiStrength​ ​and​ ​NLTK​ ​respectively.

**RQ3:** ​Kappa values obtained between different tools is slightly better compared to manual labelling but not enough to say that the tools agree with each other. Kappa values are same when WNLU is paired with either NLTK and SentiStrength. Taking into account the ARI value we can say that NLTK and WNLU agree with each​ ​other​ ​the​ ​most.

**RQ4:** ​To answer RQ4 we take intersection of tools and calculate their agreement statistics with manual labelling. Although WNLU performed best when tools were compared individually with manual labels, but the intersection of NLTK and SentiStrength agree the most with manually labelled sentiments with kappa=0.6 and​ ​ARI=0.297​ ​.

However, when intersection of all the three tools is taken and compared with manual labelling, it leads to a kappa value of 0.66 which is the highest agreement that can be achieved amongst all combinations. This could be because of drastic reduction​ ​in​ ​available​ ​sentiments​ ​because​ ​of​ ​the​ ​intersection​ ​operation.

# REFERENCES​

1. A. Murgia, P. Tourani, B. Adams, & M. Ortu, (2014, May). Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In *Proceedings of the 11th working conference on mining software repositories* (pp.​ ​262-271).​ ACM.​
2. R. Jongeling, P. Sarkar, S. Datta, & A. Serebrenik, (2017). On negative results when using sentiment analysis tools for software engineering research.​ *Empirical*​​*Software*​​*Engineering*​,​ ​1-42.
3. R. Bakeman, & J. M. Gottman, (1997). *Observing*​ *interaction: An introduction*​​*to*​​*sequential*​​*analysis*​.​ ​Cambridge​ ​university​ ​press.
4. M. Thelwall, K. Buckley, & G. Paltoglou, (2012). Sentiment strength detection for the social web. *Journal*​ *of the Association for Information Science*​​*and*​​*Technology*​, ​ *63*​(1), ​ 163-173.​
5. D. Pletea, B. Vasilescu, & A. Serebrenik, (2014, May). Security and emotion: sentiment analysis of security discussions on GitHub. In *Proceedings of the 11th working conference on mining software repositories* (pp.​ ​348-351).​ ​ACM.
6. J. Cohen, (1968). Weighted kappa: Nominal scale agreement provision for scaled​ ​disagreement​ ​or​ ​partial​ ​credit.​ *Psychological*​​*bulletin*​, ​ *70*​(4), ​ 213.​
7. A. J. Viera, & J. M. Garrett, (2005). Understanding interobserver agreement: the​ ​kappa​ ​statistic.​ *Fam*​​*Med*​,​ *37*​(5),​ ​360-363.
8. L. Hubert, & P. Arabie, (1985). Comparing partitions. ​*Journal of classification*​,​ *2*​(1), ​ 193-218.​
9. B.​Pang, & L. Lee, (2008). Opinion mining and sentiment analysis.

*Foundations*​​*and*​​*Trends®*​​*in*​​*Information*​​*Retrieval*​, ​ *2*​(1–2), ​ 1-135.​

[10]J. L. Fleiss, B. Levin, & M. C. Paik, (2013). *Statistical*​ *methods for rates and proportions*​.​ ​John​ ​Wiley​ ​&​ ​Sons.