AN ANALYSIS ON EMOJIS AND TEXT PARSING FOR SENTIMENT ANALYSIS.

[1]Pooja S. Gadani, [2]Prof. Gayatri Pandi(Jain),
[1]Master of Engineering, [2]Head of Department(PG Dept.)
M.E Computer Engineering,
L.J. Institute Of Engineering & Technology (GTU), Ahmedabad, Gujarat

Abstract: With the advance in internet technology blogs and social media generates massive volumes of data. Data relies on various ranges like small text, emojis, images and video etc. Sentiment analysis helps us to evaluate the product or service performance from user-generated contents. The Proposed technique is a combination of Senti-N-Gram technique and Emojis definition from Emojipedia. First, we classify document based on text and emoji content. For sentiment analysis of emojis, we use Emojipedia as our source for definitions. We use a rule-based approach to extract n-gram sentiment score generation. We also propose a sentiment classification methodology by using a ratio based approach based on counts of positive and negative sentences of a document. Then we compare our findings with existing approaches.

***Index Terms* – Sentiment Analysis, n-gram, Emojipedia, NLP.**

I. Introduction

In the last decade, we have seen various technological improvements and an increase in internet activities which were able to provide a positive impact on many research activities. Sentiment analysis can be defined as the computerized process of recognizing, detecting, and determining the orientation of human opinion or emotion, which is directed towards different entities [1]. Sentiment analysis is an art of identifying sentiment or emotions expressed in many languages. Individuals normally post their opinions or emotions for products, restaurants, hotels, services, movies, political issues, etc. Opinions, sentiments, and emotions can be captured using the individual's writings, facial expressions, speech and many other media [4]. Sentiment analysis aims to determine the attitude of a speaker, writer, or other subjects with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event [3].

Sentiment Analysis (Opinion Mining) is a part of Natural Language Processing (NLP) that builds systems that identify and extracts opinions and polarity like positive or negative from given text. For example, products, restaurants, hotels, services, movies, political issues, etc. extract their Opinions, sentiments, and emotions. It can be captured using the individual's writings, facial expressions, speech and many other media [4]. Sentiment analysis aims to determine the attitude of a speaker, writer, or other subjects with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event [3].

Sentiment Analysis has many applications in daily life for monitoring and analyzing public opinions regarding political issues. SA can also be used in market intelligence [5], measuring the degree of user satisfaction on products or services and improving their weaknesses [6], forecasting of price changes according to news sentiments, developing new products and services, and promoting and improving products according to customers' reviews. Being more trustworthy products and services' reviews as posted by their users compared to the vendor's reviews, many individuals rely on these reviews to make their decisions about the products, services, and other entities [7].

II. Sentiment Analysis

A. Taxonomy of sentiment analysis

We can classify Sentiment analysis into document level, sentence level and aspect level as illustrated in Fig. 1. At the document level, the main task is to extract all sentiment from the whole document, which can either be long or short. To determine the polarity of the whole document we consider the whole document as a single feature. So result here is the overall polarity of a document. In document-level sentiment analysis, we consider the whole document as one topic and decide if the overall polarity of a document is positive or negative based on some opinion words [11].

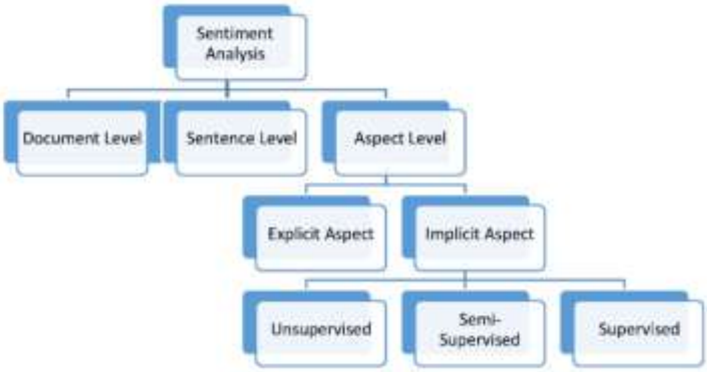
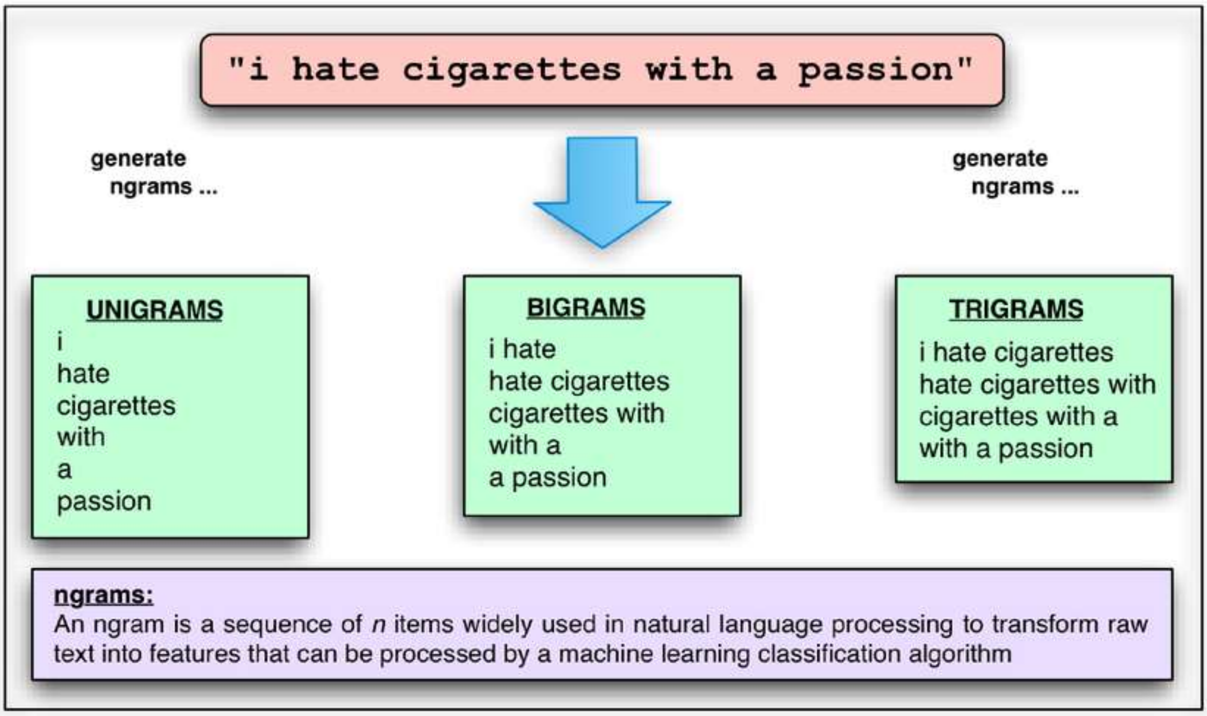


Fig. 1. Taxonomy of Sentiment Analysis[10]

At the sentence level, a process is to find overall polarity of given sentence without considering each feature a single case. At first, we identify whether the given sentence is subjective or objective and based on that we decide overall opinion of given sentence is positive or negative.

At aspect level or feature, a level is a fine-grained model in sentiment analysis. Here we determining the opinion intended by people to a specific feature (aspect) of a product, service, or any entity.[9] [11] [12] In order to perform a sentiment analysis at aspect level, we have to extract entities and their related aspects/features also known as opinion targets. Then we determine the polarity of opinions directed to a given aspect. Then we summarized the result of Sentiment Analysis task and visualized[9]. Based on Fig. 1, the extracted feature can either be explicit or implicit, where the feature is considered explicit if it is mentioned explicitly in the review sentence otherwise it is considered implicit.

III. N-Gram and Text Parsing Using NLP.



An ngram is a sequence of *n* items widely used in natural language processing to transform raw text into features that can be processed by a machine learning classification algorithm

Fig. 2. N-Gram Approaches[40]

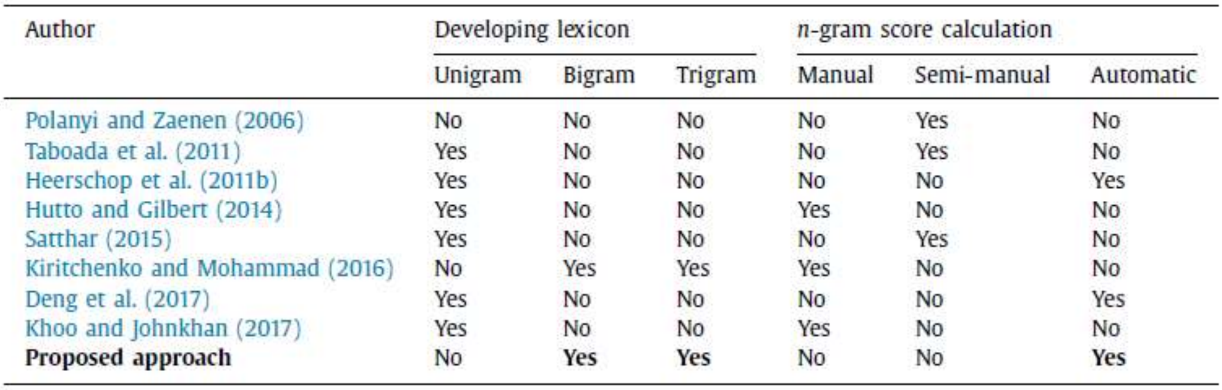
NLP - Natural Language Processing (NLP) is all about leveraging tools, techniques and algorithms to process and understand natural language-based data, which is usually unstructured like text, speech and so on[13][14][15]. We are using computers to manipulate structured data stored in form of spreadsheets and data tables. But the way we communicate with words, not in form of tables. Lots of information are hidden in the raw text of English or another human language. Natural Language Processing, or NLP, is the sub-field of AI in which we try to enable the computer to understand and process human languages.

- Steps in NLP :
  1. Lexical Analysis – It will identify and analyze the structure of words. Lexicon of a language means the collection of words and phrases in any language. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words.
  2. Syntactic Analysis (Parsing) − Analysis of words in the sentence for given grammar and arranging words in a manner that helps us to identify the relationship among the words. The sentence such as “The school goes to a boy” is rejected by English syntactic analyzer.
  3. Semantic Analysis − It draws the exact meaning or the dictionary meaning from the given text. The text is checked for its meaningfulness. We are mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as “hot ice-cream”.
  4. Discourse Integration − Meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.
  5. Pragmatic Analysis − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

IV. Analysis And Existing Work.

There are two broad approaches for calculating the sentiment of a text document: rule-based and machine learning based. The machine learning-based approaches classify the user-generated contents into positive or negative classes using some commonly used classifiers such as Naïve Bayes (NB)[16], Maximum Entropy (ME)[17], Support Vector Machine (SVM) [18] etc. The classifiers need sizable labelled datasets for training and testing [19] [20] [21] [22]. The rule-based approaches are often preferred where training datasets are hard to obtain [23] [24] [25] [26]. The rule-based approaches evaluate the sentiments using publicly available lexicons [23] [27] [24] [28].

Creating a lexicon can be manual, semi-manual or automatic. The dictionaries such as GI, WordNet, ANEW, LIWC, VADER are created using a manual method where human subjects are involved. Such manual processes are expensive in terms of both time and cost. As the name indicates semi-manual methods combines both human annotators and algorithms to build the dictionary. There have been some efforts for semi-manual (SentiWord- Net [23] and SenticNet [27]) and automatic [29] [30] [31] sentiment lexicons creation.



Author	Developing lexicon			*n*-gram score calculation		
	Unigram	Bigram	Trigram	Manual	Semi-manual	Automatic
Polanyi and Zaenen (2006)	No	No	No	No	Yes	No
Taboada et al. (2011)	Yes	No	No	No	Yes	No
Heerschop et al. (2011b)	Yes	No	No	No	No	Yes
Hutto and Gilbert (2014)	Yes	No	No	Yes	No	No
Satthar (2015)	Yes	No	No	No	Yes	No
Kiritchenko and Mohammad (2016)	No	Yes	Yes	Yes	No	No
Deng et al. (2017)	Yes	No	No	No	No	Yes
Khoo and Johnkhan (2017)	Yes	No	No	Yes	No	No
Proposed approach	No	**Yes**	**Yes**	No	No	**Yes**

Table:1- Comparison of related Approach[10]

The table shows the summary of comparison among proposed and existing approaches in terms of their pros and cons on several points. As shown in the table, first three columns differentiate these work based on whether they create a specific lexicon or not; and if at all they create whether the entries in the lexicon are in the form of unigrams, bigrams or trigrams. The last three columns show the lexicon generation procedure considering whether it is manual, semi-manual or automatic. The comparison shows, [35] [37] and [36] do not create any new lexicon but propose some semi-automatic score calculation procedure. Most of the other methods generate a unigram lexicon except [34]. They, however, use either manual or semi-manual method using human annotators except [30][38].

The basic unigram based approaches work in a straightforward way by collecting the sentiment scores associated with the words from some lexicon and add them up to find the score for the sentences [24]. Many of the recent approaches for sentiment analysis find the polarity of the text documents using sentiment unigrams score, collected from some lexicon, along with intensifiers and negations [33] [32] For example, a unigram good can be amplified with a word very or can be down toned with the word slightly. It can also be negated with use of the word not. Such n-gram methods perform better compared to the existing lexicon based unigram approach [34] [35] [36] [37].

In all the above approaches n-gram based sentiment analysis methods depend on the publicly available unigram lexicons and maintain a list of intensifier with some pre-specified scores. When a sentiment n-gram is encountered in the text the score is calculated on-line. So they are not fully automatic calculating score to create a domain-independent n-gram sentiment dictionary. Above problem addressed using Senti-N-Gram approach [39].

V. Methodology

According to the literature survey [39], there are certain limitations for performing Sentiment Analysis of text with Emojis using N-Gram approach. It is observed that n-gram text parsing approaches are more accurate and efficient compared to unigram text parsing approaches. But in the literature survey, Sentiment Analysis using N-Gram approach is done only for the input that contains text. Emojis, Punctuations and other oddly spelt words are ignored. As we know that Sentiment Analysis of Emojis are very important as they convey equal Emotion as text in which they appeared. We are going to address this problem by performing Sentiment analysis of text with Emojies using N-Gram approach.

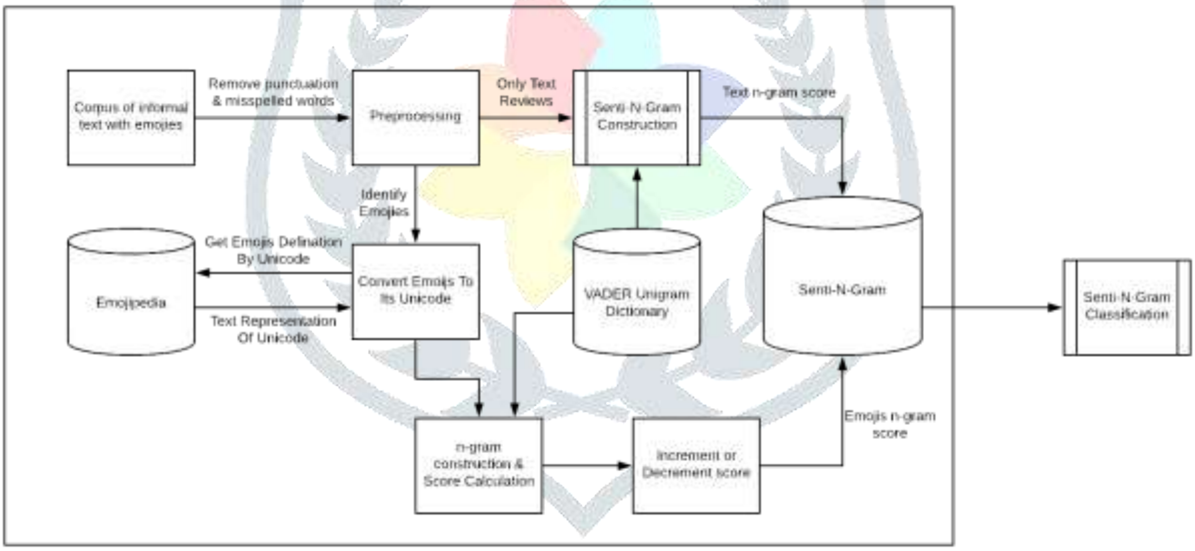

Corpus of informal text with emojies
Remove punctuation & misspelled words
Preprocessing
Only Text Reviews
Senti-N-Gram Construction
Text n-gram score
Identify Emojies
Get Emojis Defination By Unicode
Emojipedia
Convert Emojis To Its Unicode
Text Representation Of Unicode
VADER Unigram Dictionary
Senti-N-Gram
Senti-N-Gram Classification
Emojis n-gram score
n-gram construction & Score Calculation
Increment or Decrement score

Fig. 3 Block Diagram of Proposed Model

The Proposed technique is the combination of Senti-N-Gram technique and Emojis definition from Emojipedia. Preprocessing step will identify a possible occurrence of Emojis in text using regex pattern recognition. Emojipedia used to get definition and description of emojis based on its Unicode [41]. Then we construct possible n-gram for emoji text and calculate the score using VADER dictionary. Normal text Sentiment score and Emoji sentiment score are stored in Senti-n-gram database and Senti-n-gram classification process continue as in Senti-N-Gram approach [39].

VI. Conclusion And Future Work

To the best of our knowledge, the Existing literature paper gives better performance according to their training dataset but generate certain limitation like time-consuming and cost wise issues. Senti-N Gram approach[39] address all above issues but it still has a limitation of not able parse Emojis. The Proposed technique is the combination of Senti-N-Gram technique and Emojis definition from Emojipedia. The main objective to provide this approach is to eliminate the limitation presented in Senti-N-Gram approach. We fetch emoji description from Emojipedia and construct its Senti-n-gram score with some constant Increment Decrement percentage included. As we are combining sentiment score for text and emojis, we get a more accurate result for any corpus sentiment score. For future work we try to extend our work to support for sarcasm and oddly spelled words.

VII. Acknowledgment

I would like to express a deep sense of gratitude towards my guide Prof. Gayatri S. Pandi., Professor of Computer Engineering Department & HOD of PG Department for her constant encouragement and valuable suggestions. The work that we have been able to present is possible because of her timely guidance and support.

References

[1] https://www.datasciencecentral.com/profiles/blogs/text-classification-sentiment-analysis-tutorial-blog
[2] https://monkeylearn.com/sentiment-analysis
[3] https://www.knime.com/blog/sentiment-analysis-with-n-grams
[4] Mohammad Tubishat, Norisma Idris, Mohammad A.M. Abushariah " Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges"- Information Processing and Management 54 (2018).
[5] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), 25.
[6] Li, Y.-M., & Li, T.-Y. (2013). Deriving market intelligence from microblogs. Decision Support Systems, 55(1), 206–217.
[7] Kang, D., & Park, Y. (2014). Review based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, 41(4), 1041–1050
[8] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14–46
[9] Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: Comparative analysis and survey. Artificial Intelligence Review, 46(4), 459–483.
[10] Mohammad Tubishata, Norisma Idrisa, Mohammad A.M. Abushariahb, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges", Information Processing and Management 54, 2018, pp. 545-563
[11] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), 25.
[12] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. AIN Shams Engineering Journal, 5(4), 1093–1113.
[13] https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm
[14] https://towardsdatascience.com/how-to-use-natural-language-processing-to-analyze-product-reviews-17992742393c
[15] https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e
[16] McCallum, A. , Nigam, K. , et al. (1998). A comparison of event models for naive Bayes text classification. In Proceedings of the AAAI-98 workshop on learning for text categorization: 752 (pp. 41–48). Citeseer **.**
[17] Nigam, K. , Lafferty, J. , & McCallum, A. (1999). Using maximum entropy for text clas- sification. In Proceedings of the IJCAI-99 workshop on machine learning for infor- mation filtering: 1 (pp. 61–67) .
[18] Hsu, C.-W., Chang, C.-C., Lin, C.-J. et al. (2003). A practical guide to support vector classification.
[19] Dey, A. , Jenamani, M. , & Thakkar, J. J. (2017). Lexical tf-idf: An n-gram feature space for cross-domain classification of sentiment reviews. In Proceedings of the inter- national conference on pattern recognition and machine intelligence (pp. 380–386). Springer **.**
[20] Liu, S. M. , & Chen, J.-H. (2015). A multi-label classification based approach for sen- timent classification. Expert Systems with Applications, 42 (3), 1083–1093 .
[21] Pang, B. , Lee, L. , & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on empirical methods in natural language processing: 10 (pp. 79–86). Association for Computational Linguistics **.**
[22] Tripathy, A . , Agrawal, A . , & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57 , 117–126 **.**
[23] Baccianella, S. , Esuli, A. , & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lex- ical resource for sentiment analysis and opinion mining. In Proceedings of the conference on language resources and evaluation, LREC: 10 (pp. 2200–2204) .
[24] Hutto, C. J. , & Gilbert, E. (2014). Vader: A parsimonious rule-based model for senti- ment analysis of social media text. In Proceedings of the eighth international AAAI conference on weblogs and social media .
[25] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the fortieth annual meeting on association for computational linguistics (pp. 417–424). Association for Computational Linguistics **.**
[26] Hogenboom, A. , Heerschop, B. , Frasincar, F. , Kaymak, U. , & de Jong, F. (2014). Multi- -lingual support for lexicon-based sentiment analysis guided by semantics. De- cision Support Systems, 62 , 43–53 .

[27] Cambria, E. , Havasi, C. , & Hussain, A. (2012). Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In Proceedings of the Florida artificial intelligence research society conference, Flairs (pp. 202–207) .

[28] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[29] Almatarneh, S. , & Gamallo, P. (2017). Automatic construction of domain-specific sen- timent lexicons for polarity classification. In Proceedings of the international con- ference on practical applications of agents and multi-agent systems (pp. 175–182). Springer **.**

[30] Deng, S. , Sinha, A. P. , & Zhao, H. (2017). Adapting sentiment lexicons to domain-spe- cific social media texts. Decision Support Systems, 94 , 65–76 .

[31] Tan, S. , & Wu, Q. (2011). A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Systems with Applications, 38 (10), 12094–12100 .

[32] Taboada, M. , Voll, K. , & Brooke, J. (2008). Extracting sentiment as a function of dis- course structure and topicality. Technical Report **.**

[33] Heerschop, B. , van Iterson, P. , Hogenboom, A. , Frasincar, F. , & Kaymak, U. (2011c). Analyzing sentiment in a large set of web data while accounting for nega- tion. In Proceedings of the advances in intelligent web mastering–3 (pp. 195–205). Springer .

[34] Kiritchenko, S. , & Mohammad, S. M. (2016). Happy accident: A sentiment composi- tion lexicon for opposing polarity phrases. In Proceedings of tenth edition of the language resources and evaluation conference (LREC) . Portoroz, Slovenia .

[35] Polanyi, L. , & Zaenen, A. (2006). Contextual valence shifters. In Computing attitude and affect in text: Theory and applications (pp. 1–10). Springer .

[36] Satthar, F. S. (2015). Modelling SO-CAL in an inheritance-based sentiment analy- sis framework. In Proceedings of the OASIcs—OpenAccess Series in informatics: 49 . Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik .

[37] Taboada, M. , Brooke, J. , Tofiloski, M. , Voll, K. , & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37 (2), 267–307 **.**

[38] Heerschop, B. , Hogenboom, A. , & Frasincar, F. (2011b). Sentiment lexicon creation from lexical resources. In Proceedings of the international conference on business information systems (pp. 185–196). Springer

[39] Atanu Dey , Mamata Jenamani , Jitesh J. Thakkar , “Senti-N-Gram : An n -gram lexicon for sentiment analysis”, Expert Systems With Applications 103, 2018, pp. 92-105

[40] https://www.researchgate.net/figure/N-gram-text representation_fig4_256290162

[41] Milagros Fernández-Gavilanes ∗, Jonathan Juncal-Martínez , Silvia García-Méndez , Enrique Costa-Montenegro, FranciscoJavier González-Castaño, “Creating emoji lexica from unsupervised sentiment analysis of their descriptions”, Expert Systems With Applications 103 , 2018, pp. 74-91