****

**Faculty of Information Technology**

**Information Systems**

**Sentiment Analysis on Amazon Reviews**

**Final Project**

**Report**

**Done by: Slamova Gaukhar,**

**Nurbekova Tamiris**

**Checked by: Jongtae Park**

**Almaty, 2021**

**Motivation and Objective**

With the development of the network, the approach of client service changed. Many people are using the Internet for online shopping and other purposes. Organizations need to investigate clients' behavior to recommend them the best product or to review products' shortcomings, to find these we utilize clients' opinions with Sentiment Analysis. Amazon is one of the biggest companies, and there is a lot of data. We use Amazon product reviews to find the best machine learning method for Sentiment Analysis.

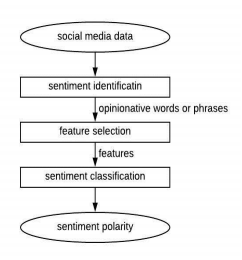
**Related Work and Originality**

The Sentiment Analysis is the computational investigation of individuals' feelings, frames of mind, and feelings toward a substance. The element can speak about any people, occasion, or theme, we can collect them by surveys. This recognizes the feeling communicated in a book at that point and investigates it. Consequently, feeling examination aims to discover sentiments, distinguish the notions they express, and afterward arrange their extremity [1]. Sentiment Analysis in many cases viewed as an arrangement procedure. There are 3 primary characterization levels in sentiment analysis: document level, sentence level, and aspect level. Document-level plans to characterize a conclusion archive as communicating a positive or negative supposition or estimation while Sentence-level intends to arrange to feel clarified in each sentence. The first step is to distinguish whether the sentence is abstract or goal, it can be an emotional sentence, so sentence-level supposition examination will decide if the sentence communicates positive or negative sentiments. Be that as it may, there is no essential contrast among document and sentence level orders since sentences are simply short archives [1]. The perspective level of sentiment analysis means to group the opinions as for the particular parts of substances. The initial step is to recognize and to see the substances and their viewpoints. The reviewers can give various conclusions for various parts of a similar substance like this sentence "The battery of telephone is not working very well". This research handles the initial two types of opinion examination, which means that sentiment analysis utilizes the NLP content investigation and computational procedures to computerize the extraction or arrangement of review (opinion) from assessment surveys. Investigation of these sentiments has spread crosswise over numerous fields of study, for example, consumer data, marketing, books, applications, sites, and social media. Opinion examination turns into a hot region in basic leadership [2]. A huge number of clients and companies rely upon online feeling surveys and questionnaires. The main objective of inquiring supposition is to break down the surveys and look at the scores of reviews or opinions. This investigation is separated into numerous levels [3]: document level [4], sentence-level [5], word/term level [6], or aspect level [7]. The succession procedure of sentiment analysis is the assessment and recognition of the sentiment polarity [8]. The authors in the papers [9],[10],[11], have highlighted the conclusion challenges. The datasets fully utilized in evaluation examination are a significant or main issue in this field. The primary pieces of information are from the surveys, so these surveys are imperative to the businessmen as they can take business choices as indicated by the investigation after-effects of clients' sentiments about their product. The audit sources are fundamentally surveyed like in Amazon company. [1] Sentiment examination isn't just applied to item audits however can likewise be applied to financial exchanges, news stories, or political discussions. In political discussions, we could make sense of individuals' sentiments on specific political race up-and-comers or ideological groups. Interpersonal companies’ buyers are viewed as a generally excellent wellspring of data since individuals share and talk about their assessments [12]. They are also utilized as information sources in the conclusion investigation process, so the feature choice procedures are given in subtleties in fig.1 alongside their related articles alluding to some beginning references. The Sentiment Classification methods are talked about with more subtleties representing related articles and starting references too.[13]

**Design Architecture and Detailed Algorithm**

The main approach that is discussed in various papers is the reviews given by customers to the products like in Amazon company, or commenting about a movie, etc. To understand these product reviews we mainly sentiment analysis. How the text is pre-processed, the feature selection methods used, and what are the classifier models needed are discussed in further sections. These reviews are converted into simple English language which is easily understandable and classify whether the review is positive, negative, or neutral. The feature extraction methods and classifiers are the main steps in the study of sentiment analysis.

Figure 1. Process flow



***Pre-processing***

Raw reviews are parsed from the Amazon website, so generally, the result is a noisy dataset. This is due to the casual nature of people’s usage of the internet, mostly they don’t pay attention to grammar and spelling. Reviews from the Amazon website have certain special characteristics such as emoticons, URLs, etc. which have to be suitably extracted. Therefore, raw review data has to be normalized to create a dataset that can be easily learned by various classifiers. In this research work, we have applied several pre-processing steps to standardize the dataset and to reduce its size. We first did some general pre-processing on reviews as follows.

* Convert the review to lower case.
* Replace two or more dots with space.
* Strip spaces and quotes.
* Replace two or more spaces with a single space.

Url

Users often share hyperlinks to other web pages in their reviews. Url is not important to this work, it may lead to very sparse features. Therefore, we replace all the URLs in reviews with the word URL. The regular expression used to match URLs is ((www\.[\S]+)|(HTTPS?://[\S]+)).

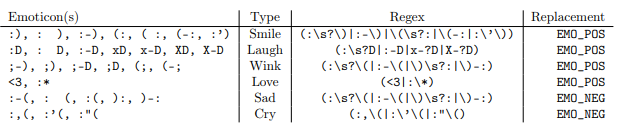
User Mention

Users or reviewers often mention other users in their reviews by using @handle, so we have replaced all user mentions with the word user\_mention. The regular expression used to match user mention is @[\S]+.

Emoticons

Users often use several different emoticons in their reviews to convey different emotions. It is impossible to exhaustively match all the different emoticons used on websites as the number is ever increasing. In this paper, we tried to match some common and frequent emoticons, then we replaced the matched emoticons with either emo\_pos or emo\_neg depending on whether it is a positive or negative emotion. A list of all emoticons matched by our method is given in table 1.

Table 1: List of emoticons



Hashtag

Hashtags are unspaced phrases prefixed by the hash symbol (#) which is frequently used by users to mention trending topics. We deleted hashtags from words, an example, #hello to hello. Here, the regular expression used for hashtags is #(\S+).

Re-review

Re-review are reviews that have already been sent by someone else and are shared by other users. Re-reviews began with the letters RR, in our case we removed RR from the reviews as it is not an important feature for text classification. The regular expression used to match Re-review is \brt\b. After applying review level pre-processing, we processed individual words of reviews as follows.

* Strip any punctuation [’"?!,.():;] from the word.
* Convert two or more letter repetitions to valid letters. Some people send reviews like I am sooooo happpppy adding multiple characters to emphasize certain words. This is done to handle such reviews by converting them to I am soo happy.
* Remove - and ’. It used to handle words like a t-shirt and aren’t by converting them to tshirt and arent.
* Check whether the word is valid. We identified valid words as words that began with an alphabet with successive characters being alphabets, numbers, or one of a dot (.) and underscore(\_).

Porter Stemmer

Finally, we used Porter Stemmer in the NLTK library which is a process for removing the common morphological and inflexional endings from words in the English language and mainly used as part of a term normalization process.

Lemmatization

In the project, we used the ready-made WordNetLemmatizer tool of NLTK library. Then we used, stop-words of the English language.

**Machine Learning Models:**

Naive Bayes is a simple model which can be used for classification of text, in our case reviews. In this model, the class cˆ is assigned to a tweet t, where cˆ = argmax c P(c|t) P(c|t) ∝ P(c)Yn i=1 P(fi |c) In the formula above, fi represents the i-th feature of total n features. P(c) and P(fi |c) can be obtained through maximum likelihood estimates.

SVM, also known as support vector machines, is a non-probabilistic binary linear classifier. For a training set of points (xi , yi) where x is the feature vector and y is the class, we want to find the 7 maximum-margin hyperplane that divides the points with yi = 1 and yi = −1.

The equation of the hyperplane is as follow w · x − b = 0 We want to maximize the margin, denoted by γ, as follows

max w,γ γ, s.t.∀i, γ ≤ yi(w · xi + b)

in order to separate the points well.

**Results and Performance Evaluation**

Figure 2. Performance Evaluation of Naive Bayes

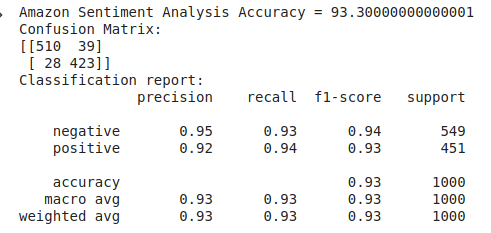
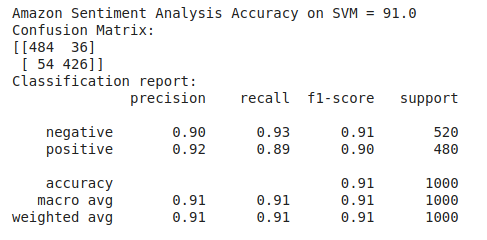


Figure 3. Performance Evaluation of SVM



From the above methods, our best model is the Multinomial Naive Bayes. In statistics, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models. There are different variants of the Naive Bayes (NB) classifier in sklearn. We used Multinomial Naive Bayes as a big word but it just means many classes/categories/bins/boxes that need to be classified. We got the 93% accuracy, which is very good.

**Conclusion and Future Work**

In this project, for the Sentiment Analysis, we mainly focused on preprocessing and word embedding. With the help of the data cleaning, we got a pretty good result, which may not be achieved without preprocessing. For the word embedding, we used lemmatization and found frequent words, and created vocabulary.

In future work, we hope to use more machine learning algorithms to test our hypothesis that with preprocessing every model can give good results.

**References**

1. Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative Study of Machine Learning Approaches for Amazon Reviews. Procedia Computer Science,132, 1552- 1561. doi:10.1016/j.procs.2018.05.119
2. Chalothom, Tawunrat, and Jeremy Ellman. ―Simple Approaches of Sentiment Analysis via Ensemble Learning.‖ Lecture Notes in Electrical Engineering Information Science and Applications, 2015, pp. 631– 639., doi:10.1007/978-3-662-46578-3\_74.
3. “A Study on Sentiment Analysis: Methods and Tools”. International Journal of Science and Research (IJSR), vol. 4, no. 12, May 2015, pp. 287–292., doi:10.21275/v4i12.nov151832.
4. Yin, Yichun, et al. “Document-Level Multi-Aspect Sentiment Classification as Machine Comprehension”. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, doi:10.18653/v1/d17-1217.
5. Farra, Noura, et al. “Sentence-Level and Document-Level Sentiment Mining for Arabic Texts.” 2010 IEEE International Conference on Data Mining Workshops, 2010, doi:10.1109/icdmw.2010.95.
6. E. Nikos, L. Angeliki , P . Georgios, C. Konstantinos ELS: a word-level method for entity-level sentiment analysis WIMS’11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics (2011) doi:10.1145/1988688.1988703.
7. Li, Chunping, et al. “Modeling Sentiment and Aspect Using Syntax: A Topic Model Approach.” Proceedings of The 5th IIAE International Conference on Industrial Application Engineering 2017, 2017, doi:10.12792/icisip2017.036.
8. Khan, Khairullah, et al. “Mining Opinion Components from Unstructured Reviews: A Review.” Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 3, 2014, pp. 258–275., doi:10.1016/j.jksuci.2014.03.009.
9. Pozzi, F.a., et al. “Challenges of Sentiment Analysis in Social Networks.” Sentiment Analysis in Social Networks, 2017, pp. 1–11., doi:10.1016/b978-0-12-804412-4.00001- 2.)
10. Liu, Bing, and Lei Zhang. “A Survey of Opinion Mining and Sentiment Analysis.” Mining Text Data, 2012, pp. 415– 463., doi:10.1007/978-1-4614-3223-4\_13.
11. Mukherjee, Arjun, et al. “Spotting Fake Reviewer Groups in Consumer Reviews.” Proceedings of the 21st International Conference on World Wide Web - WWW 12, 2012, doi:10.1145/2187836.2187863.
12. J, A., & G, J. (2018). Sentiment Classification of Tweets with Non-Language Features. Procedia Computer Science,143, 426-433. doi:10.1016/j.procs.2018.10.414
13. Jain, Vardhan, Kandasamy. “Sentiment Analysis Of Product Reviews – A

Survey”. International Journal of Scientific & Technology Research, volume 8, issue 12, December 2019. ISSN 2277-8616.

**Roles of Members**

**Slamova Gaukhar -** Research, Preprocessing, Word Embedding

**Nurbekova Tamiris -** Research, Multinomial NB, SVM