# DATA 698 Literature Review

**Topic: Student Counselling Software**

**Irene Jacob**

## Literature Review

Social media sites like Twitter, Facebook, blogs, and many online forums produce large amount of unstructured data. These large amounts of unstructured data are classified as positive, negative, or neutral. Positive sentiments are those which contain appreciation for other's tweets, movies, products, etc while negative sentiments are those which contain bad words or dissatisfactory comments on any product, movie, event, etc. Whereas neutral sentiments are neither positive nor negative sentiment. Surveys were conducted using various approaches of clustering with respect to sentiment analysis and this presents a way to find relationships between the tweets based on polarity and subjectivity. For efficient data analysis lexicon-based approach was found to be more efficient (Deshapnde et al., 2019; Gupta et al., 2017).

Machine learning is the study of algorithms that can learn from and make predictions on data. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. A naïve bayes algorithm is very easy to build up and mainly used for a large set of data. On the other hand, a neural network has two phases, which are training and testing. In training phase, the positive and negative comments are trained and assigned weights. The main purpose of training phase is to create the dictionary of positive comments. In the next phase testing is done based on the weighted dictionary. The artificial neural network is trained with labelled data to produce meaningful output. This process by which neural networks learn from labelled data is called as back propagation (Siddharth et al., 2018).

An approach that can be generalized to any domain was presented by Khattak et al. (2020), and they have used it to extract healthcare knowledge from publicly available tweets, providing recommendations for diabetes. They have used Java and other open APIs to create an application which amalgamates the data curation service, knowledge extraction service, user profile building service, and filter engine into the proposed recommendation system. By applying seed list-based classification and sentiment analysis, the system was able to recommend personalized diabetes-related tweets to users. To overcome redundancy problems and formatting issues, Google Refine is used. To calculate the accuracy of the proposed system, they used seed list for diabetes for tweet filtration.

There are many social networking sites available these days, and Faizan (2019) has used twitter data in his paper to carry out his research. Twitter is in a lot of fame at present because of its specific format of writing. Few of the characteristics of tweets are that tweets are short messages consisting of a maximum of 140 characters. Approximately 1.2 billion tweets are posted daily, and these are posted by a wide variety of people. Thus, the topics discussed are also variant and could almost include any topic starting from politics to mobile products etc.

Experiments for two classification tasks were carried out by Agarwal et al. (2011), namely, positive versus negative and positive versus negative versus neutral. For each of the classification tasks three models were presented along with the results for two combinations of these models. For the unigram plus Senti-features model, a feature analysis was carried out to gain insight on the type of features that added most value to the model. All the experiments used Support Vector Machines (SVM) and reported an average 5-fold cross-validation test results. Tree kernel and feature based models were also investigated and it was discovered that both these models outperform the unigram baseline. It was revealed that the most important features are those that combine the prior polarity of words and their parts-of-speech tags. A conclusion was made that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres.

Machine learning techniques like the Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) produced the greatest precision, especially when multiple features were included. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive. Machine learning algorithms, such as The Naive Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram model were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve a classification accuracy of approximately 85%. In general, it was expected that ensemble Twitter sentiment-analysis methods would perform better than supervised machine learning algorithms, as they combined multiple classifiers and occasionally various feature models. However, hybrid methods also performed well and obtained reasonable classification accuracy scores (Alsaeedi & Khan, 2019).

A dataset of different Arabic dialects, which consists over 151,500 tweets/comments, was collected and automatically labelled. The NB, LR, ME, PA, RR, SVM, MNB, Ada-Boost BNB, and SGD classifiers were used to extract and discover the polarity of a given tweet. A 10-fold cross validation was utilized to divide the data into a separated training set and testing set. The best evaluation metric values were achieved by PA and RR using unigram, bigram, or trigram is 99.96% (Gamal et al., 2019). This work is concerned with SA in Arabic textual content.

## References


1. Deshapnde, P., Joshi, P., Pawar, P., Madekar, D., & Salunke, P. (2019). *A Survey On: Sentiment Analysis framework of Twitter data Using Classification*. Retrieved 30 March 2022.
2. SIDDHARTH, S., DARSINI, R., & SUJITHRA, D. (2018). *SENTIMENT ANALYSIS ON TWITTER DATA USING MACHINE LEARNING ALGORITHMS IN PYTHON*. Retrieved 30 March 2022.
3. Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). *Study Of Twitter Sentiment Analysis Using Machine Learning Algorithms On Python*. Retrieved 30 March 2022.
4. Khattak, A., Batool, R., Satti, F., Hussain, J., Khan, W., Khan, A., & Hayat, B. (2020). *Tweets Classification And Sentiment Analysis For Personalized Tweets Recommendation*. Retrieved 30 March 2022.
5. Faizan, F. (2019). *Twitter Sentiment Analysis*. Retrieved 30 March 2022.
6. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment Analysis Of Twitter Data*. Retrieved 30 March 2022.
7. Alsaeedi, A., & Khan, M. (2019). *A Study On Sentiment Analysis Techniques Of Twitter Data*. Retrieved 30 March 2022.
8. Gamal, D., Alfonse, M., El-Horbarty, E., & Salem, A. (2019). *Implementation Of Machine Learning Algorithms In Arabic Sentiment Analysis Using N-Gram Feature*. Retrieved 30 March 2022.
9. X. Chen, M. Vorvoreanu and K. Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences," in IEEE Transactions on Learning Technologies, vol. 7, no. 3, pp. 246-259, July-Sept. 2014, doi: 10.1109/TLT.2013.2296520.
10. Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014, Pages 1093-1113, ISSN 2090-4479, https://doi.org/10.1016/j.asej.2014.04.011.
11. C. Kaur and A. Sharma, "Social Issues Sentiment Analysis using Python," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-6, doi: 10.1109/ICCCS49678.2020.9277251.
12. Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.