Final Paper Proposal

1. Members of the final project team and their proposed roles in the project.

The project team includes the following members: Kirill Dolgin, Betty Kao, Almazhan Kapan.
The responsibilities will be divided as follows:

Almazhan Kapan - Writing, Programming & Software Development, Theoretical Issues, Evaluation.
Betty Kao - Programming & Software Development, Theoretical Issues, Evaluation.
Kirill Dolgin - Evaluation, Theoretical Issues.

Theoretical issues are topics related to computer science, math and/or linguistics. Responsibilities are divided based on the skills of the members and their preference. Almazhan has some experience with research and writing. Betty and Almazhan have some experience with using CNN and related neural networks. All members have some experience with theoretical knowledge.

2. Problem statement and motivation.

Currently, there are more than 4.6 billion Internet users worldwide with many of them actively sharing their opinion on topics ranging from products and services to political events (Statista, 2021). Collecting and analyzing these opinions is the main objective of sentiment analysis and is useful for a variety of purposes from brand management to election result predictions. In particular, sentiment analysis is a field of study in NLP that detects polarity or emotion at levels of words, phrases, sentences or documents. Commonly, the text element is classified as positive, negative or neutral (Hogenboom, 2015).

Earlier research in sentiment analysis has treated natural language text as a 'bag-of-words' and proposed both supervised and unsupervised solutions. While supervised solutions can achieve higher accuracy with more training, their classification results might not be intuitive and might not perform well on the non-trained data (emoticon). Due to these reasons, lexicon-based methods, particularly ones based on SVM classifiers and linguistic features, still remain popular (Heerschop, 2011).

However, one remaining challenge even for lexicon-based sentiment analysis is that sentiment of a phrase depends not just on what words are used (lexicon), but also on how these words are used (Bal, 2011). Even in face-to-face conversations, visual cues such as frowning or smiling can outweigh the impact of words in detecting a person's mood or attitude (Ferreri, 2018). Given this context, emoticons or emojis, which have been overlooked in earlier sentiment analysis studies, might be particularly useful in detecting public sentiment. For example, a seemingly objective sentence 'Classes start tomorrow' can drastically change its sentiment based on whether a negative or positive emoticon is used along with the sentence i.e. 'Classes start tomorrow :) ' or 'Classes start tomorrow :-('.

This project focuses on sentiment analysis of tweets using SVM classifiers and rich linguistic features, building on the previous lexicon-based methods (Mohammad, 2013) and additionally introduces new linguistic features (placing importance on emojis and emoticons) to detect the sentiment of a tweet (Hogenboom, 2015). This project is inspired by the SemEval 2017 Task 4 Subtask A where the participants are asked to identify the sentiment of the tweet and classify it as positive, negative or neutral. SemEval is one of the leading forums in the field of sentiment analysis that allows researchers to conduct sentiment analysis using Twitter data at different granularity since 2013 (Nakov et. al 2013).

3. Evaluation measures and dataset.

Datasets. We use the testing dataset provided by the forum organizers and is annotated using CrowdFlower. The collected tweets are based on popular topics and events trending on Twitter using a Twitter related named entity recognition system (Ritter et al., 2011). The topics include geopolitical entities (e.g. country names), named entities (e.g. Barack Obama) and other entities ('Western media'). Twitter API was used to download tweets and additional filtering applied e.g. only topics with more than 100 tweets remained and near-duplicates were removed (Rosenthal et al, 2016). For training and development purposes, we use datasets from the previous SemEval years from 2013 to 2016 and use the 30% of the test dataset provided by the organizers. Topics for the test dataset differ from the topics for the training and development datasets.

Evaluation measures. SemEval tasks that involve Twitter sentiment analysis have been traditionally treated as a 'single-label multi-class' task where each tweet can belong only to one class: Positive, Negative or Neutral. For the baseline model, the project will follow evaluation guidelines defined in the SemEval 2017 Task 4, Subtask A with some slight modifications.

In particular, instead of calculating the 'average recall' value which was the primary measure for SemEval 2017 Task 4, we will use the 'macro f-score', which was used in previous SemEval years and is based on f-scores for the 'Positive' and 'Negative' classes of tweets.

As mentioned above, our goal is to enrich the model proposed by the NRC Canada team (Mohammad et.al) and the latter was trained and tested with the 'macro f-score' as the primary evaluation metric. Hence, we include 'macro f-score' to successfully compare these two models.

Moreover, the training set for our model is collected from training and testing sets from the previous SemEval years, which also used 'macro f-score'.Moreover, the task organizers noted that 'average recall' is chosen since it provides a better consistency for Subtask B. However, in our project we focus only on Subtask A, therefore, replacing 'average recall' is not necessarily significant while using 'macro f-score' provides certain benefits mentioned above.

The 'macro f-score' denoted as F_1^{PN} will be computed as below:

$$F_1^{PN} = (F_1^{P} + F_1^{N}) / 2$$

F_1^{P} refers to the f-score for the Positive class of tweets and F_1^{N} refers to the f-score for the Negative class of tweets. F_1^{P} will be computed as below:

$$F_1^{P} = (2\text{prec}^{P} * \text{recall}^{P}) / (\text{prec}^{P} + \text{recall}^{P})$$

In this formula prec^{P} and recall^{P} denote precision and recall for the Positive class and can be computed as follows:

$$\text{prec}^{P} = \frac{PP}{PP+PL+PN} \qquad \text{recall}^{P} = \frac{PP}{PP+LP+NP}$$

PP, PU, PN, UP, NP refer to the entries in the confusion matrix below specified in SemEval 2016 Task 4. For each cell AB, the number of tweets labeled by the system as A and the gold standard labelled as B. P refers to Positive, L refers to Neutral and N refers to Negative classes of tweets (Nakov, 2016).

		Gold Standard		
		Positive	Neutral	Negative
Predicted	Positive	PP	PU	PN
	Neutral	UP	UU	UN
	Negative	NP	NU	NN

Similarly to f-score for the Positive class, we will calculate f-score for Negative class and then compute the 'macro-average' f-score.

4. Academic papers.

Some of the articles that we will use for our project are listed below. The full list is available in the References section.

1. Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 502–518. https://www.aclweb.org/anthology/S17-2088.

This article provides an overview of the SemEval 2017 Task 4: Sentiment Analysis in Twitter. Particularly, the paper describes common goals for each subtask, discusses datasets, annotation process, performance evaluation measures and also mentions top participating teams and their chosen methods to solve a subtask. This article is particularly valuable for our project since our system project attempts to solve Subtask A of SemEval 2017 Task 4 and information about evaluation, common methods is helpful to assess the performance of the system and also view examples of other projects with the same goal.

2. Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2014. NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, pages 443–447. https://www.aclweb.org/anthology/S14-2077.

This model outlined in this article solves the task of sentiment analysis on Twitter data, particularly, SemEval 2013 Task 4: Sentiment Analysis in Twitter. The SVM system outlined in this article was scored first in SemEval 2013 and is generally considered one of the most successful models based on linguistic features. This article is especially useful for our project as the system mentioned in this article is based on rich lexical and semantic features and our project also uses lexicon and feature based approach. Our project will refer to some of the features described in this paper (POS, Negation etc.) and also add new linguistic features.

3. Alexander Hogenboom, Danella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. In *Journal of Web Engineering,* 14(1–2), Rinton Press Incorporated, Paramus, NJ, pages 022-040. https://dl.acm.org/doi/10.5555/2871254.2871257

This article analyzes how using sentiment conveyed by emoticons can serve to improve the performance of sentiment analysis models. This paper is useful for our project since it provides justification for using polarity of emoticons and also shows how to integrate this knowledge into lexicon and linguistic feature based projects. The article also provides an example for creating an emoticon based lexicon, which we plan to do for our project

4. Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 2096–2101. https://www.aclweb.org/anthology/N19-1214.

This article demonstrates the importance of using emojis and emoticons in sentiment analysis tasks and provides several strategies for processing emojis, particularly, using embeddings and using their natural language descriptions. This paper is useful for our project as it provides common justification for using emojis and emoticons in sentiment analysis. Moreover, the model outlined in the paper is built with neural networks, particularly, using a stack of two BiLSTMs with attention networks. Thus, the article will be particularly useful for stage 3 in our project as we also start using attention networks with RNN.

5. Matthias Hagen, Martin Potthast, Michel B¨uchner, and Benno Stein. 2015. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval*

2015). Association for Computational Linguistics, Denver, Colorado, pages 582–589. https://www.aclweb.org/anthology/S15-2097.

The model outlined in this article solves the task of Sentiment Analysis and received a top score for the SemEval 2015 Task 10: Sentiment Analysis in Twitter project. The project reproduces four Twitter sentiment classification approaches that participated in prior SemEval editions. This article is very useful as its system refers to the model produced by the NRC Canada team (mentioned above), which allows to compare the performance of the features across datasets and also employs rich lexicon based and semantic features, which can potentially provide inspiration for our project.
The

5. Project Strategy.

The problem our system will solve is detection of sentiment of the tweet as positive, negative or neutral, which closely follows SemEval 2017 Task 4 Subtask A. This particular task has attracted the greatest number of teams since the SemEval forum was first started and different supervised and unsupervised approaches have been proposed.

Supervised models usually incorporate rich linguistic features and among them, the model proposed by Mohammad et al. for the SemEval 2013 Task 4 is found as one of the most successful models. This model is built using SVM classifiers and incorporates lexicon and semantic features such as n-grams, POS tags, clusters, punctuations, emojis.

For the baseline version of the project, we will use some of the features proposed by team NRC Canada (Mohammad et. al) and add additional linguistic features and emoji based sentiment analysis rules.
We train the Support Vector Machine (SVM) in the same manner as the NRC Canada model in order to better evaluate the impact of individual features (both original and new) on the dataset.SVM is considered an effective algorithm for text classification tasks (Mohammad, 2013).

For every tweet in the training dataset, we extract feature vectors and then train in the SVM. Then, for every test tweet, we extract feature vectors and then predict the output using SVM.
The following features will be used for our project:

Lexicon based features. The following Lexicons will be used for the project: NRC-Emotion, NRC-Sentiment140, NRCHashtag (Mohammad et al., 2013), MPQA Lexicon (Wilson et al., 2005) and the Bing Liu Lexicon (Hu and Liu, 2004). 4 features will be extracted from these lexicons: total count of tokens with positive scores, total score of the tweet, maximal word score in tweet and score of the last token with positive score. In addition, unlike the NRC Canada team, we will create our own Emoticon lexicon and assign scores to emoticons between -1 and 1; we will use emoticon scores as an additional feature (details below).

Word N-grams. We tokenize the training tweets and create a dictionary with all unary and binary words. We extract the features using these dictionaries. For every feature, if it appears in the tweet, the feature will be equal to how many times the feature appeared. Otherwise, the feature will be equal to 0.

Adjectives. From analysis of related submissions for the task, using N-grams might produce features that are too sparse and vector size is too large (Abreu, 2017). In this case, we might decide to focus only on adjectives since they are most descriptive. Particularly, we might only use unigrams and create a dictionary of adjectives.

Negation. We use the number of negated contexts as a feature. Negated context refers to a segment of a tweet that starts with a negation word (such as not, shouldn't, etc.) and ends with one of the following punctuation marks: ',', '.', ':', ';', '!', '?' (Pang et. al 2002). Negated context will affect N-gram and lexicon based features. The list of negation words is derived from Christopher Potts' tutorial on sentiment.
Part of Speech tags. We tokenize the data and apply part of speech tagging using either NLTK or CMU Pos Tagger. Specific elements such as URL or hashtags will be normalized. The number of times each part-of-speech tag occurs will be used as a feature.

Emoticons. We check presence or absence of positive and negative emoticons in the tweet and use the sum of all scores as a feature. We also might use polarity of the last emoticon in the tweet as an additional feature.
We will refer to the emoticon based sentiment analysis pipeline proposed by Hogenboom et. al
and our own emoticon dataset.

For the base model (stage 1), we will train SVM with the following features: Lexicon based features (excluding the Emoticon lexicon), Word N-grams, Negation, Part of Speech Tags. For our prediction model (stage 2), we will train SVM with additional features such as Emoticons and Adjectives. Given time and resources, we will also train the given data with a model based on neural networks (stage 3), since the top scoring teams for SemEval 2017 employed deep learning. In particular, we will train data using RNN with attention networks.
Compared to CNN, RNN is generally more preferable for NLP tasks since the former does not retain information about the word order.

6. A collaboration plan.

The project tasks can be divided into independent subsets of tasks.

1. Preprocessing. Team member 1: Preprocess the data (remove duplicate tweets, mentions, replace URL, etc).
2. Create an Emoticon based lexicon: all members.
3. Building different features for the model. Provisionary plan can be as follows: based on the ‘gold’ data (i) team member 1 will create Lexicon based features; (ii) team member 2 will create Word N-grams, POS, Negation based features; (iii) team member 3 will create Emoticon based features.
4. Train the data with SVM. Provisionary plan can be as follows: (i) team member 1 will create training code for the features that Team member 1 has built (make sure both teammates have the same parameters on training code for easier later collaboration!); (ii) team member 2 will create training code for the features that Team member 2 has built; (iii) team member 3 will create training code for the features that Team member 3 has built
5. Evaluation. Calculate the ‘macro f-score’ for the training outputs from all team members.
6. Merge the individual programs together and potentially create unit tests and test the program.
7. Train the data using different combinations of features (Features from all teammates) and calculate the ‘macro f-score’ as outlined in the Evaluation section. All teammates will train data on different combinations of features. For example, one team member can train the data using all Emoticon based features, while the other member will train the data without Emoticon based features, to see if the f-score changes.
8. Similarly to training the data, all members will test the data on the various combinations of features.

References

Matthias Hagen, Martin Potthast, Michel B¨uchner, and Benno Stein. 2015. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 582–589. https://www.aclweb.org/anthology/S15-2097.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 502–518. https://www.aclweb.org/anthology/S17-2088.

Sabino Miranda-Jiménez, Mario Graff, Eric Sadit Tellez, and Daniela Moctezuma. 2017. INGEOTEC at SemEval 2017 Task 4: A B4MSA Ensemble based on Genetic Programming for Twitter Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 771–776. https://www.aclweb.org/anthology/S17-2130.

Mohammed Jabreel and Antonio Moreno. 2017. SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich Set of Features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 694–699. https://www.aclweb.org/anthology/S17-2115.

José Abreu, Iván Castro, Claudia Martínez, Sebastián Oliva, and Yoan Gutiérrez. 2017. UCSC-NLP at SemEval-2017 Task 4: Sense n-grams for Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 807–811. https://www.aclweb.org/anthology/S17-2136.

Alexander Hogenboom, Danella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. 2015. Exploiting emoticons in polarity classification of text. In *Journal of Web Engineering,* 14(1–2), Rinton Press Incorporated, Paramus, NJ, pages 022-040. https://dl.acm.org/doi/10.5555/2871254.2871257

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 321–327. https://www.aclweb.org/anthology/S13-2053.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, Toronto, Ontario, Canada, pages 1678–1684.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1615–1625. https://www.aclweb.org/anthology/D17-1169.

Abu Awal Md Shoeb, Shahab Raji, and Gerard De Melo. 2019. EmoTag – Towards an Emotion-Based Analysis of Emojis. In *12th International Conference on Recent Advances in Natural Language Processing, RANLP 2019*. Incoma Ltd, Varna, Bulgaria, pages 1094-1103.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, pages 24-33. https://www.aclweb.org/anthology/S18-1003.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 2096–2101. https://www.aclweb.org/anthology/N19-1214.

Yinxia Lou, Yue Zhang, Fei Li, Tao Qian, and Donghong Ji. 2020. Emoji-Based Sentiment Analysis Using Attention Networks. In *ACM Trans. Asian Low-Resour. Lang. Inf. Process*. Association for Computing Machinery, New York, NY, USA, pages 2375-4699. https://dl.acm.org/doi/abs/10.1145/3389035.

AltexSoft. 2018. Sentiment Analysis: Types, Tools, and Use Cases. [online] Available at: <https://www.altexsoft.com/blog/business/sentiment-analysis-types-tools-and-use-cases/> [Accessed 13 April 2021].

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 123–131. https://www.aclweb.org/anthology/W16-5415.

Statista. 2021. *Internet users in the world 2021*. [online] Available at: <https://www.statista.com/statistics/617136/digital-population-worldwide/> [Accessed 13 April 2021].

B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. 2011. Polarity Analysis of Texts using Discourse Structure. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)*. Association for Computing Machinery, New York, NY, USA, pages 1061-1070. https://dl.acm.org/doi/10.1145/2063576.2063730.

D. Bal, M. Bal, A. van Bunningen, A. Hogenboom, F. Hogenboom, and F. Frasincar. 2011. Sentiment Analysis with a Multilingual Pipeline. In *12th International Conference on Web Information System Engineering (WISE 2011)*. volume 6997 of Lecture Notes in Computer Science, pages 129-142. https://www.semanticscholar.org/paper/Sentiment-Analysis-with-a-Multilingual-Pipeline-Bal-Bal/76843afcfcb4b5880808f8540cd2d29105f433c7.

Ferretti V, Papaleo F. Frasincar. 2019. Understanding others: Emotion recognition in humans and other animals. Genes Brain Behav 18(1). https://onlinelibrary.wiley.com/doi/full/10.1111/gbb.12544