

2018

A Framework to Understand Emoji Meaning: Similarity and Sense Disambiguation of Emoji using EmojiNet

Sanjaya Wijeratne
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all



Part of the [Computer Engineering Commons](#), and the [Computer Sciences Commons](#)

Repository Citation

Wijeratne, Sanjaya, "A Framework to Understand Emoji Meaning: Similarity and Sense Disambiguation of Emoji using EmojiNet" (2018). *Browse all Theses and Dissertations*. 2227.
https://corescholar.libraries.wright.edu/etd_all/2227

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

A Framework to Understand Emoji Meaning: Similarity and Sense Disambiguation of Emoji using EmojiNet

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

By

Sanjaya Wijeratne
B.Sc. in Information Technology, University of Moratuwa, Sri Lanka, 2009.

2018
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

November 19, 2018

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Sanjaya Wijeratne ENTITLED A Framework to Understand Emoji Meaning: Similarity and Sense Disambiguation of Emoji using EmojiNet BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Amit Sheth, Ph.D.
Dissertation Director

Michael Raymer, Ph.D.
Director, Computer Science and Engineering
Ph.D. Program

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Committee on Final Examination

Amit Sheth, Ph.D.

Derek Doran, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Wenbo Wang, Ph.D.

ABSTRACT

Wijeratne, Sanjaya. Ph.D., Department of Computer Science and Engineering, Wright State University, 2018. A Framework to Understand Emoji Meaning: Similarity and Sense Disambiguation of Emoji using EmojiNet.

Pictographs, commonly referred to as ‘emoji’, have become a popular way to enhance electronic communications. They are an important component of the language used in social media. With their introduction in the late 1990’s, emoji have been widely used to enhance the sentiment, emotion, and sarcasm expressed in social media messages. They are equally popular across many social media sites including Facebook, Instagram, and Twitter. In 2015, Instagram reported that nearly half of the photo comments posted on Instagram contain emoji, and in the same year, Twitter reported that the ‘face with tears of joy’ emoji has been tweeted 6.6 billion times. As of 2017, Facebook and Facebook Messenger processed over 60 million and 6 billion messages with emoji per day, respectively. Emogi, an Internet marketing firm, reports that over 92% of all online users have used emoji at least once. Creators of the SwiftKey Keyboard for mobile devices report that they process 6 billion messages per day that contain emoji. Moreover, business organizations have adopted and now accept the use of emoji in professional communication. For example, Appboy, an Internet marketing company, reports that there has been a 777% year-over-year increase and 20% month-over-month increase in emoji usage for marketing campaigns by business organizations in 2016. These statistics leave little doubt that emoji are a significant and important aspect of electronic communication across the world.

The ability to automatically process and interpret text fused with emoji will be essential as society embraces emoji as a standard form of online communication. In the same way that natural language is processed with sophisticated machine learning techniques and technologies for many important applications, including text similarity and word sense disambiguation, emoji should also be amenable

to such analysis. Yet the pictorial nature of emoji, the fact that the same emoji may be used in different contexts to express different meanings, and that emoji are used in different cultures over the world which can interpret emoji differently, make it especially difficult to apply traditional Natural Language Processing (NLP) techniques to analyze them. Indeed, emoji were developed organically with no overt/explicit semantics assigned to them. This contributed to their flexible usage but also lead to ambiguity. Thus, similar to words, emoji can take on different meanings depending on context and part-of-speech (POS). Polysemy in emoji complicates determination of emoji similarity and emoji sense disambiguation. However, having access to machine-readable sense repositories that are specifically designed to capture emoji meaning can play a vital role in representing, contextually disambiguating, and converting pictorial forms of emoji into text, thereby leveraging and generalizing NLP techniques for processing richer medium of communication.

This dissertation presents the creation of EmojiNet, the largest machine-readable emoji sense inventory that links Unicode emoji representations to their English meanings extracted from the Web. EmojiNet consists of (i) 12,904 sense labels over 2,389 emoji, which were extracted from reliable online web sources and linked to machine-readable sense definitions seen in BabelNet; (ii) context words associated with each emoji sense, which are inferred through word embedding models trained over Google News and Twitter message corpora for each emoji sense definition; and (iii) recognizing discrepancies in the presentation of emoji on different platforms and specification of the most likely platform-based emoji sense for a selected set of emoji. It then discusses the application of emoji meanings extracted from EmojiNet to solve novel downstream applications including emoji similarity and emoji sense disambiguation. To address the problem of emoji similarity, first, it presents a comprehensive analysis of the semantic similarity of emoji through emoji embedding models learned over emoji meanings in EmojiNet. Using emoji descriptions, emoji sense labels, and emoji sense definitions, and with different training corpora obtained from Twitter and Google News, multiple embedding models are learned to measure emoji similarity. Using a benchmark sentiment analysis dataset, it further shows that incorporating emoji meanings in EmojiNet into embedding

models can improve the accuracy of sentiment analysis tasks by $\sim 9\%$. To address the problem of emoji sense disambiguation, it uses word embedding models learned over Twitter and Google News corpora and shows that word embeddings models can be used to improve the accuracy of emoji sense disambiguation tasks. The EmojiNet framework, its RESTful web services, and other benchmarking datasets created as part of this dissertation are publicly released at <http://emojinet.knoesis.org/>.

Contents

1	Introduction	1
1.1	Focus of the Dissertation	6
1.1.1	EmojiNet	6
1.1.2	Solving Emoji Understanding Tasks using EmojiNet	7
1.1.3	Emoji Similarity	8
1.1.4	Emoji Sense Disambiguation	9
1.2	Dissertation Organization	11
2	Background and Related Work	13
2.1	Overview	13
2.2	Emoticons, Emoji, and Other Pictographs	13
2.3	Semiotics, Writing Systems, and Emoji	15
2.4	Emoji Understanding and Building Emoji Sense Dictionaries	17
2.5	Emoji Similarity	21
2.6	Emoji Sense Disambiguation	23
2.7	Summary	25
3	EmojiNet: Building a Machine-Readable Emoji Sense Inventory	26
3.1	Overview	26
3.2	Emoji Modeling and Dataset Creation	27

3.3	Open Resources Used in EmojiNet	28
3.3.1	The Unicode Consortium Emoji List	28
3.3.2	Emojipedia	29
3.3.3	The Emoji Dictionary	29
3.3.4	BabelNet	29
3.4	Resource Integration	30
3.4.1	Linking Resources based on the Unicode Code Points	30
3.4.2	Linking Resources based on the Images	31
3.4.3	Extracting Sense Labels	31
3.4.4	Extracting and Linking with BabelNet Senses	33
3.5	Enhancing EmojiNet for Analysis Tasks	35
3.5.1	Adding Word Embeddings to EmojiNet	35
3.5.2	Adding Platform-specific Meanings to EmojiNet	36
3.6	EmojiNet Web application and REST API	37
3.7	Dataset Evaluation	40
3.7.1	Resource Integration Evaluation	40
3.7.2	Sense Assignment Evaluation	41
3.8	EmojiNet at Work - An Experiment on Calculating Emoji Similarity	43
3.9	Summary	45
4	Applications of EmojiNet – Emoji Similarity	47
4.1	Overview	47
4.2	Representation of Emoji Meaning	48
4.2.1	Emoji Description (<i>Sense_Desc.</i>)	48
4.2.2	Emoji Sense Labels (<i>Sense_Label</i>)	48
4.2.3	Emoji Sense Definitions (<i>Sense_Def.</i>)	49

4.2.4	Learning the Emoji Embedding Models	49
4.3	Ground Truth Data Curation	52
4.3.1	Emoji Pair Selection	54
4.3.2	Human Annotation Task	55
4.3.3	Annotation Evaluation	56
4.4	Evaluating Emoji Embedding Models	60
4.5	Emoji Embeddings at Work	62
4.6	Summary	64
5	Applications of EmojiNet – Emoji Sense Disambiguation	65
5.1	Overview	65
5.2	Proposed Approach	66
5.3	Summary	69
6	Conclusion and Future Work	70
6.1	Overview	70
6.2	EmojiNet: Building a Machine-Readable Emoji Sense Inventory	70
6.3	Emoji Similarity Calculation	71
6.4	Emoji Sense Disambiguation	71
6.5	Future Work	72
6.5.1	Building a Machine-Readable Emoji Sense Inventory	72
6.5.2	Emoji Similarity Calculation	73
6.5.3	Emoji Sense Disambiguation	73
6.5.4	Emoji Prediction	74
6.6	Going Forward	75
6.7	Summary	76
	Appendices	77

A Platform-specific Emoji in EmojiNet	78
A.1 Overview	78
B EmojiNet REST API Calls	80
B.1 Overview	80
Bibliography	85

List of Figures

1.1	Emoji Usage in Social Media with Multiple Senses.	3
2.1	Emoji Usage Growth on Instagram. Image Source – http://bit.ly/2QVaKTS	14
3.1	Construction of Emoji Representation in EmojiNet.	30
3.2	Using The Unicode Consortium and The Emoji Dictionary for Sense Label Filtering.	33
3.3	Assigning BabelNet Sense Definitions.	34
3.4	EmojiNet Web Application at http://emojinet.knoesis.org/	38
3.5	Emoji Sense Distribution.	40
3.6	Emoji Clusters using Emoji Sense Overlap.	44
4.1	Learning Emoji Embedding Models using Word Vectors.	50
4.2	Emoji Co-Occurrence Frequency Graph.	53
4.3	A Screen shot of the Web Application Used for the Annotation Task.	56
4.4	Top-5 Emoji Pairs with Highest Inter-annotator Agreement for Each Ordinal Value from 0 to 4.	58
4.5	Distribution of the Mean of User Ratings.	59
5.1	Emoji Usage in Social Media with Multiple Senses.	66

List of Tables

3.1	Nonuple Representation of an Emoji.	27
3.2	EmojiNet Statistics.	39
3.3	Word Sense Disambiguation Statistics.	42
3.4	Selected Emoji Sense Clusters in EmoTwi50.	45
3.5	Ten Most Similar Emoji Pairs Based on Jaccard Similarity.	46
4.1	Spearman’s Rank Correlation Results.	61
4.2	Accuracy of the Sentiment Analysis task using Emoji Embeddings.	63
5.1	Top 10 Emoji based on the Emoji Sense Disambiguation Accuracy (in % values).	68
A.1	Platform-specific Emoji in EmojiNet.	79
B.1	Get Emoji Information.	81
B.2	Get Emoji Images.	81
B.3	Get Noun Meanings for Emoji.	82
B.4	Get Verb Meanings for Emoji.	82
B.5	Get Adjectives Meanings for Emoji.	83
B.6	Get Twitter Word Embeddings for Emoji.	83
B.7	Get Google News Word Embeddings for Emoji.Emoji (Twitter)	84

ACKNOWLEDGEMENTS

I joined Kno.e.sis Center, Wright State University in Fall 2011, and my stay here has been a very interesting one. I've met amazingly talented people throughout my stay at Kno.e.sis who inspired me to strive for the best in everything I do. I'd like to take this opportunity to express my heartfelt gratitude for all who helped me during this journey.

First and foremost, I would like to thank my dissertation advisor Dr. Amit P. Sheth. The opportunity that he gave me to work under his guidance played a huge part in my life as a graduate student. He was very keen and patient to see how I grow as a research student, being supportive in every success and failure of mine. He encouraged me every time I failed by guiding me to correct my mistakes. He challenged me when he noticed that I haven't reached to my full potential and cheered me up until I reach there. He went out of the way to help me to find internships and made sure I'm prepared to solve real-world problems. He showed me the importance of non-technical skills such as effective communication, networking, presentation, and teamwork, and provided an ecosystem to master them. He always encouraged internal and external collaborations, and thanks to him, I got to work closely with many professors from various research backgrounds that helped me to learn the skill set required to collaborate in diverse environments. I always admire his vision and knack for selecting the next best research problem to solve. I've greatly benefited by these qualities of his as I work with him. He never let me worry about anything other than research and supported my studies intellectually and financially. I wouldn't be able to achieve any of my accomplishments during my stay at Kno.e.sis if it wasn't for his continuous support and guidance. Therefore, I'm deeply grateful for him for everything he has done for me and I'm very fortunate to work under his guidance.

I would also like to express my sincere gratitude to the rest of my dissertation committee, Dr.

Derek Doran, Dr. Krishnaprasad Thirunarayan (a.k.a. T. K. Prasad), and Dr. Wenbo Wang. Dr. Doran and I started collaborating when Dr. Sheth suggested me to work with him for an invited paper. Since then, I closely worked with Dr. Doran for many years. I've learned many things from Dr. Doran, especially, when it comes to paper writing. He would give detailed, line-by-line reviews and revise my papers until every author is certain that we submit the best paper we can write. I cannot overstate the value of Dr. Doran's inputs and his effect on my academic writing abilities. He was always available for me to discuss matters related to my research. During our discussions, Dr. Doran would push my boundaries to make sure that I do and achieve my level best. I'm extremely thankful for him for his time and effort to help me grow as a graduate student and my sincere gratitude and appreciation go to him.

Dr. Prasad was very inspirational to me while I was at Kno.e.sis. I worked very closely with Dr. Prasad when I was working on my very first research paper and I was blown away by the knowledge he posses on diverse areas. That lead me to seek his guidance whenever I'm stuck at a research problem. He has helped me to improve many of my research works by providing valuable feedback on my papers and presentations. He would ask very important questions whenever I present my work during our weekly group meetings, which certainly helped to better shape my work. Dr. Prasad's presentations on physics and science are inspirational and helped me to maintain my interest in those fields. I loved the discussions we had on Vidur's and Neeti's accomplishments and I'm extremely thankful to Dr. Prasad for everything he has done for me.

I've known Dr. Wenbo Wang since the first day of my graduate school and he was very helpful for me since then. Dr. Wang possess many qualities that I admire. Especially, I was amazed by his perseverance during graduate school and it inspired me to keep focused and work towards my goals even in situations where I failed. He shared what worked best for him during graduate school and taught me about prioritizing my work and managing my time. He encouraged me to collaborate with Lakshika, with whom I co-authored many papers, later. He encouraged me to finish my studies on time and he spent several evenings with me preparing me for my Ph.D. proposal and dissertation

presentations. He asked important questions during the presentations which helped to better shape my work. I express my sincere gratitude and appreciation for everything Dr. Wang has done for me.

I was very fortunate to collaborate with many of my colleagues at Kno.e.sis center and it gave me a great opportunity to learn from them. My first mentor was Ajith Ranabahu. Ajith was a dear friend who showed me how research is done. He was very patient with me and extremely helpful. Not only that, he and his wife Dharshi helped me to settle down in the USA. They were my unofficial godparents while they were here. Then, I was mentored by Prateek Jain. With Prateek, I got to work on Linked Open Data-related projects. Prateek was very helpful in providing feedback on my work and helped me to work towards the project goals. Then, I was mentored by Pavan Kapanipathi and it was the first time that I worked with social media data. We worked on exciting problems related to temporal entity ranking and continuous semantics. Apart from many things I learned from Pavan, he helped me to find my first internship at The Insight Centre for Data Analytics, National University of Galway, Ireland. I'm extremely grateful for these mentors for their time, efforts, and helping me to grow as a researcher. I also got to collaborate with Kalpa Gunaratne, Lakshika Balasuriya, Francois Lamy, and Anurag Illendula. It was fun working with them. I also got to collaborate with faculty from diverse research backgrounds. I worked with Dr. Robert Carlson and Dr. Raminta Daniulaityte from The Center for Interventions, Treatment, and Addictions Research (CITAR) and Dr. Jack Dustin from the Center for Service-Learning and Civic Engagement at Wright State University. I learned many skills from them. Especially, I learned how to explain computer science problems and concepts to people from a non-computer science background, which is an important skill to possess when working with clients. I'm extremely lucky to work with them and I thank them for everything they have done for me.

During my stay at Kno.e.sis, I got to meet a truly diverse group of friends, from whom I learned many things. We traveled to places together and had cookouts, game nights, movie nights, and many more fun-filled activities when we were away from work. Especially, I would like to thank Ajith, Dharshi, Sujan, Sarasi, and Kalpa for all the good times we had. We've become very good

friends over the years and continue to support each other even after graduating from Kno.e.sis, for which, I'm truly thankful to all of you. I would also like to thank Prateek, Cory, Vinh, Ashutosh, Pavan, Revathy, Delroy, Wenbo, Lu, Pramod, Hemant, Ashutosh, Vinh, Shereyansh, Topher, Mary, Raghava, Alan, Jeremy, Koneru, Swapnil, Surendra, Pavan Kalyan, Roopteja, Pranav, Adharsh, Siva, Nishita, Venkatesh, Ankitha, Vishnu, Gaurish, Manas, Utkarshani, SoonJye, Dipesh, Shruti, Usha, Swati, Amelie, Matt, Jace, Scott, Joy, Sagar, Vaikunth, Revathy V., Ruwan, Thilini, Sreeram, Hima, Hussein, Rochelle, Melissa, and Joey. It was fun working with you all. You've proof read my papers, helped me with creating evaluating datasets, and participated in evaluation studies, thus, I'm extremely thankful for the support I received. I would also like to thank Dr. Tanvi Banerjee and Dr. William Romine for their great support. I will miss hanging out with them over weekends.

I'd also like to thank Dr. Bahareh Heravi, Dr. Rajaraman Kanagasabai, and Dr. Shonali Krishnaswamy whom I met during my internships. I got to work with Dr. Bahareh Heravi when I interned at The Insight Centre for Data Analytics, National University of Galway, Ireland. Dr. Bahareh gave me the freedom to work on research problems that interest me. Then, I interned at the Institute for Infocomm Research - A*Star, Singapore, where I was mentored by Dr. Rajaraman Kanagasabai and Dr. Shonali Krishnaswamy. There, I got to work on exciting machine learning problems and Dr. Raja and Dr. Shonali were very supportive and inspirational. I'm very grateful for those two organizations for providing me opportunities to intern and for my mentors, for their immense help and support towards making me a better researcher.

Finally, I'm extremely grateful to my family. Especially, my wife, Lakshika, who was with me during the thick and thin. She was a great colleague at work who encouraged me to thrive for the best, a collaborator with whom I had many papers with and a loving wife at home who took care of me. I'm lucky to have her by my side throughout this journey and we have made heaps of unforgettable memories during this journey together. I'm thankful to my parents, Rohana and Dhammika, and my sister Pavithri, and my friends and relatives who encouraged me throughout my stay at the grad school. It was not easy for me to stay away from them, however, technology helped

us to stay close even though we were miles apart. I know my parents are really proud of what I've achieved in grad school and I'm glad if I could put a smile back onto their faces after leaving them to pursue my dreams for the past seven years. I'd also like to thank my in-laws, Sarath, Augusta, Hemal, Kasun, Thiloka, Lakmini, Pasan, and Lakmali for the immense support. I'm also grateful for my cousins, Dheep, Tania, Dhinu, Kaumal, Sashi, Sadish, and Buddhika, for the immense support I received. I'm also grateful for the Sri Lankan community in Dayton, Ohio. They were our family away from family in Sri Lanka and they made sure that we never felt away from home. Especially, I would like to thank Gamini, Manori, Vipul, Asha, Anil, Ujitha, Pandula, Asela, Nirojan, Shiral, and Renu for their unconditional love and support.

I would also like to thank Mrs. Tonya Davis. Tonya took care of all non-academic issues for me, thus, she made it easier for me to focus on my research. I also thank Mr. Jibril Ikharo, who helped me with proofreading my papers. I would also like to thank Mrs. Jennifer Limoli and the other faculty and staff members of the Computer Science Department at Wright State University. I'm grateful to Nicole Selken, the designer of The Emoji Dictionary and Jeremy Burge, the founder of Emojipedia for giving us permission to use their web resources for our research. Finally, I would like to acknowledge the funding agencies. The work carried out in this dissertation was supported by the National Science Foundation (NSF) award: CNS-1513721: "Context-Aware Harassment Detection on Social Media", the National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02: "Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use", and the National Institutes of Mental Health (NIMH) award: 1R01MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression". Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the NSF, NIDA, or NIMH.

Dedicated to

my grandfather, late Mr. R. M. W. Wijeratne

my parents, Rohana and Dhammika Wijeratne

my wife Lakshika and my sister Pavithri

1

Introduction

Thesis Statement: Machine-readable emoji sense repositories can be created and used to enable substantially better understanding of the emoji meaning in text contexts. This is useful for improving the performance of downstream applications such as emoji sense disambiguation and calculating emoji similarity.

With the rise of social media, pictographs, better known as ‘emoji’, have become an extremely popular form of communication. Their popularity may be explained by the typical short text format of social media, with emoji able to express rich content in a single character. With their introduction in the late 1990’s, emoji have been widely used to enhance the sentiment [Novak et al. 2015], emotion [Wood and Ruder 2016], and sarcasm expressed in social media messages [Joshi et al. 2017; Felbo et al. 2017]. They are taking over non-standard orthographies used in social media such as slang terms [Dimson 2015] and emoticons [Pavalanathan and Eisenstein 2016]. For example, Instagram reported that the users of their platform prefer emoji over slang terms when posting photo comments. They reported that the slang terms such as “xoxo”, “omg”, “lol”, and “hehe” are often replaced by their corresponding emoji equivalents such as 😂, ❤️, and 🙌 [Dimson 2015]. Pavalanathan *et al.* reported that Twitter users are also shifting to emoji [Pavalanathan and Eisenstein 2015; 2016]. All major social media platforms have reported that they are seeing an increase in the emoji usage in

their social media platforms. In 2018, Twitter reported that they processed 250 million emoji per month¹. Facebook reported that over 700 million messages with emoji are shared on their platform every day while over 900 million emoji are sent in Facebook Messenger without any text content every day². Studies have also shown that emoji use in social media increases user engagement³. Emoji are also a powerful way to express emotions or a hard to write, subtle notion effectively [Kelly and Watts 2015]. For example, emoji are used by many Internet users, irrespective of their age. Emogi, an Internet marketing firm reports that over 92% of all online users have used emoji and emoji usage is not simply a millennial fad, as over 65% of frequent and 28% of occasional Internet users over the age of 35 use emoji⁴. Emoji are heavily used in business communications too. For example, reports suggest that there has been a 777% year-over-year increase and 20% month-over-month increase in emoji usage for marketing campaigns in 2016⁵. Emoji have already started to blending into the Internet application markets as well. Recently, YouTube started supporting emoji-based music search and retrieval through YouTubeMusic service⁶. Many domain name service providers now support emoji domain names, which are unique Website names that consist of emoji characters⁷. Emoji have also played a role in social issues such as identifying hate crime witnesses online. For example, the ‘eye in speech bubble’ emoji  has been extensively used to widen the reach and engagement of youth in “I Am A Witness” anti-bullying campaign spearheaded by The Advertising Council (Ad Council). Above statistics and use-cases attest for the importance of emoji in online and business communications.

As analysis and modeling of written text by Natural Language Processing (NLP) techniques have enabled important advances such as machine translation [Weaver 1955], word sense disambiguation [Navigli 2009], and search [Guha et al. 2003], and the transfer of such methods (or development

¹<https://memeburn.com/2018/07/twitter-most-tweeted-emoji/>

²<https://blog.emojipedia.org/facebook-reveals-most-and-least-used-emojis/>

³<http://blog.emojics.com/emojis-increase-user-engagement/>

⁴<https://goo.gl/C5ioV0>

⁵<https://goo.gl/ttxyP1>

⁶<https://dailym.ai/2QkFUio>

⁷<https://gizmodo.com/emoji-domains-are-the-future-maybe-1823319626>





					
Sense	Example	Sense	Example	Sense	Example
Laugh (noun)	I can't stop laughing 😂	Kill (verb)	He tried to kill one of my brothers last year. 🖱️🖱️	Costly (Adjective)	Can't buy class la 💰
Happy (noun)	Got all A's but 1 😂😄	Shot (noun)	Oooooooh shots fired! 🖱️🖱️	Work hard (noun)	Up early on the grind 💰
Funny (Adjective)	Central Intelligence was damn hilarious! 😂	Anger (noun)	Why this the only emotion I know to show anger? 🖱️	Money (noun)	Earn money when one register /w ur link 💰

Figure 1.1: Emoji Usage in Social Media with Multiple Senses.

of new methods) over emoji is only beginning to be explored [Wijeratne et al. 2017a]. Only recently have there been efforts to mimic standard NLP techniques used for machine translation, word sense disambiguation, and search into the realm of emoji. However, adopting traditional NLP systems for emoji understanding is hindered due to many reasons including, (i) the graphical nature of emoji that requires additional methods to map pictographic characters to text (or Unicode), (ii) the ambiguity of emoji or the variations of emoji meaning based on how it is being used and who uses it, and (iii) emoji are used in all languages over the world which make it especially difficult to develop traditional NLP techniques [Miller et al. 2016; Barbieri et al. 2016]. Indeed, when emoji were first introduced, they were defined with no rigid semantics attached, which allowed people to develop their own use and interpretation⁸. Thus, similar to words, emoji can take on different meanings depending on context and part-of-speech (POS) [Wijeratne et al. 2016]. For example, consider the three emoji 😂, 🖱️, and 💰 and their use in multiple tweets in Figure 1.1. Depending on context, we see that each of these emoji can take on wildly different meanings. People use the 😂 emoji to mean laughter, happiness, and humor; the 🖱️ emoji to discuss killings, shootings or anger; and the 💰 emoji to express that something is expensive, working hard to earn money or simply to refer to money.




Knowing the meaning of an emoji can significantly enhance applications that study, analyze,


⁸<https://goo.gl/ztqjC2>

and summarize electronic communications. For example, rather than stripping away emoji in a preprocessing step, sentiment analysis application reported in [Novak et al. 2015] uses emoji to improve its sentiment score. For example, consider the two tweets T_1 and T_2 given below. The content of the two tweets is the same, except for T_2 , which has an additional  emoji at the end⁹. Both tweets express a negative sentiment.


T_1 : I love you and now you're just gone

T_2 : I love you and now you're just gone 

Both T_1 and T_2 contain strong sentiment-bearing words such as **love** and **gone**, which are associated with positive and negative sentiment, respectively. A simple sentiment analysis algorithm that counts only for the sentiment-bearing words to determine the sentiment score of a tweet would find it difficult to determine the correct sentiment of T_1 and T_2 . However, if you consider the presence of  emoji in T_2 as a sentiment-bearing term, the same simple sentiment analysis algorithm can be used to come up with the correct sentiment of T_2 . The  emoji is generally associated with negative sentiment¹⁰ and the presence of  in T_2 clearly indicates the negative sentiment associated with it.

Emoji can be helpful to detect sarcasm in social media posts. For example, consider the two tweets T_3 and T_4 given below. The content of the two tweets is the same, except for T_4 , which has an additional  emoji at the end¹¹. Both tweets express sarcasm.

T_3 : I love how you never reply back..

T_4 : I love how you never reply back.. 

Both T_3 and T_4 contain sentiment-bearing words such as **love** and **never**, which are associated with positive and negative sentiment, respectively. Moreover, those sentiment-bearing words are used to express opposite sentiments about the same sentiment target (i.e., the task of “replying back”).

⁹The tweet examples are adopted from [Felbo et al. 2017]

¹⁰http://kt.ijs.si/data/Emoji_sentiment_ranking/

¹¹The tweet examples are adopted from [Felbo et al. 2017]

Having opposite sentiments towards a common sentiment target is considered an important feature to determine sarcastic comments expressed online [González-Ibáñez et al. 2011; Joshi et al. 2017]. The 😞 emoji is generally associated with negative sentiment¹² and the presence of 😞 in T_4 further strengthens the negative sentiment associated with the sentiment target. Therefore, considering emoji as features could help computer programs to automatically detect sarcastic messages posted on social media.

Similar to how sentiment associated with emoji improves sentiment analysis, knowing the meaning of an emoji can also help to further improve the performance of sentiment analysis applications. A good example for this scenario would be the 😂 emoji, where people use it to describe both happiness (using senses such as laugh, joy) and sadness (using senses such as cry, tear). Knowing under which sense the 😂 emoji is being used could help to understand its sentiment better. But to enable this, a system needs to understand the particular meaning or *sense* of the emoji in a particular instance. However, prior research reports that no resources have been readily made available for this task [Miller et al. 2016]. This calls for the need of a machine-readable *sense inventory for emoji* that can provide information such as: (i) the plausible part-of-speech tags (PoS tags) for a particular use of emoji; (ii) the definition of an emoji and the senses it is used in; (iii) example uses of emoji for each sense; and (iv) links of emoji senses to other inventories or knowledge bases such as BabelNet or Wikipedia. Current research on emoji analysis has been limited to emoji-based sentiment analysis [Novak et al. 2015], emoji-based emotion analysis [Wang et al. 2012], representation learning for emoji [Barbieri et al. 2016; Eisner et al. 2016], and applications that analyze emoji [Jiang et al. 2017; Santhanam et al. 2018; Balasuriya et al. 2016; Wijeratne et al. 2016] etc. Having access to machine-readable sense repositories that are specifically designed to capture emoji meaning can play a vital role in representing, contextually disambiguating, and converting pictorial forms of emoji into text, thereby leveraging and generalizing NLP techniques for processing richer medium of communication.

¹²http://kt.ijs.si/data/Emoji_sentiment_ranking/

1.1 Focus of the Dissertation

This dissertation presents a framework which provides emoji meanings that can be used to solve emoji understanding tasks. The proposed framework, named EmojiNet, addresses the challenges with emoji interpretation identified earlier by providing machine-readable emoji meanings. Such machine-readable emoji meanings can be further used to solve downstream emoji understanding applications such as emoji similarity calculation and emoji sense disambiguation. Thus, this dissertation also examines how EmojiNet can be used to solve those emoji understanding tasks. The work presented in this dissertation makes significant contributions to the field of emoji research. Firstly, it presents a framework consists of machine-readable emoji meanings that can be used to learn emoji meanings and representations. Given the miscommunications in emoji use due to varying interpretations by different cultures and geographies, the framework presented here provides essential tools for machines to learn emoji representations. Secondly, it demonstrates how traditional NLP algorithms can be used to analyze emoji pictographic characters, enabling the processing of multi-modal data (i.e., emoji and text) using traditional NLP algorithms. Finally, it demonstrates the impact of the technologies presented in the dissertation on downstream applications such as sentiment analysis. However, those technological contributions can also benefit other research areas such as emotion analysis and emoji prediction. Below, we discuss the contributions of this dissertation, in brief, emphasizing how EmojiNet can be a key enabler to solve emoji understanding tasks.

1.1.1 EmojiNet

EmojiNet is an open service and public API that provides machine-readable emoji meanings. The service enables researchers and practitioners to query an extensive database of emoji senses and enables the potential integration of emoji with practical and theoretical NLP analyses. EmojiNet attaches 12,904 sense definitions to over 2,389 emoji, along with data about the relevance of a sense to the platform it is read on for a selected set of emoji. The set of sense definitions extracted

from BabelNet for each emoji are strengthened with context words learned from word embedding models from corpora of Google News articles and Twitter messages. This dissertation details the architecture of EmojiNet, including its integration with other web resources. It then discusses the extent of the EmojiNet emoji sense database and the format and metadata stored in it. It also gives an evaluation of the quality of emoji pictograph mapping, the quality of the BabelNet sense extraction process, and a qualitative user study using Amazon Mechanical Turk to determine the overall quality of the sense matchings to emoji and the platform it may be rendered on for a set of 40 emoji.

1.1.2 Solving Emoji Understanding Tasks using EmojiNet

Solving emoji understanding tasks such as emoji similarity calculation and emoji sense disambiguation would require a machine to learn representations of emoji meanings. There are several learning methods available such as unsupervised learning and supervised learning that can be used to learn such representations. However, these methods need access to large datasets of emoji meanings in order to learn. Especially, creating labeled datasets (i.e., training data) to solve problems such as emoji sense disambiguation can be extremely challenging due to the number of emoji meanings available for each emoji. For example, EmojiNet lists more than 30 meanings for 😂 emoji¹³, which is one of the most popular emoji across social media platforms. Creating a labeled dataset to disambiguate the meaning of 😂 would require human annotators to label social media posts carrying all meanings associated with 😂, which is a challenging task. Past research has shown that having access to knowledgebases can improve the performance of unsupervised and supervised learning methods [Sheth et al. 2017]. Thus, emoji meaning knowledgebases can be a key enabler in learning representations to solve emoji similarity calculation and emoji sense disambiguation tasks. Therefore, we also examine the performance of applying EmojiNet to solve emoji understanding tasks in this dissertation.

¹³<http://bit.ly/2PRNeSj>

1.1.3 Emoji Similarity

When studying emoji understanding, emoji similarity is considered a primary problem because foundational to many emoji analysis tasks, there should be a way to measure *similarity*. Having such measure is important for many applications including: (i) corpus searching, where documents (or a query) contains emoji symbols [Cappallo et al. 2015]; (ii) sentiment analysis [Barbieri et al. 2016; Eisner et al. 2016], where emoji sentiment lexicons [Novak et al. 2015] are known to improve the performance; and (iii) interface design, mainly in optimizing mobile phone keyboards [Pohl et al. 2017]. In fact, as of 2017, the poor design of emoji keyboards for mobile devices may be relatable to the reader: there are 2,823¹⁴ emoji supported by the Unicode Consortium, yet listing and searching through all of them on a mobile keyboard is a time consuming task. Grouping similar emoji together could lead to optimized emoji keyboard designs for mobile devices [Pohl et al. 2017].

The notion of the similarity of two emoji is very broad. One can imagine a similarity measure based on the pixel similarity of emoji pictographs, yet this may not be useful since the pictorial representation of an emoji varies by mobile and computer platform [Miller et al. 2016; Tigwell and Flatla 2016; Cramer et al. 2016]. Two similar looking pictographs may also correspond to emoji with radically different senses (e.g. twelve thirty 🕒 and six o'clock 🕒, raised hand 🙌 and raised back of hand 🙏, octopus 🐙 and squid 🐙 etc.) [Wijeratne et al. 2016; 2017a]. Instead, we are interested in measuring the *semantic* similarity of emoji, such that the measure reflects the *likeness of their meaning, interpretation or intended use*. Understanding the semantics of emoji requires access to a repository of emoji meanings and interpretations. EmojiNet offers free and open access to an aggregation of such meanings and interpretations (called senses) collected from major emoji databases on the Internet (e.g. The Unicode Consortium, The Emoji Dictionary, and Emojipedia).

A collection of emoji sense definitions can enable a semantics-based measure of similarity through vector word embeddings. Word embeddings are a powerful and proven way [Mikolov et al. 2013] to measure word similarity based on their meaning. They have been widely used in semantic similarity

¹⁴Statistics as of December 2018. For latest information, please visit <https://emojipedia.org/stats/>

tasks [Hill et al. ; Huang et al. 2012; Camacho-Collados et al. 2015] and empirically shown to improve the performance of word similarity tasks when used with proper parameter settings [Levy et al. 2015]. Word vectors also provide a convenient way of comparing them across each other. Thus, representing the emoji meanings using word embedding models can be used to generate word vectors that encode emoji meanings, which we call **emoji embedding models**.

This dissertation presents a comprehensive study on measuring the semantic similarity of emoji using emoji embedding models. As a first step to calculate emoji similarity, the machine-readable emoji meanings are extracted from EmojiNet to model the meaning of an emoji. Using pre-trained word embedding models learned over a Twitter dataset of 110 million tweets and a Google News text corpus of 100 billion words, the extracted emoji meanings are encoded to obtain emoji embedding models. To create a gold standard dataset for evaluating how well the emoji embeddings measure similarity, a group of ten human annotators who are knowledgeable about emoji were asked to manually rate the similarity of 508 pairs of emoji. This dataset of human annotations, which is called ‘EmoSim508’, is made publicly available with this dissertation. We evaluate the emoji embeddings by first establishing that the similarity measured by our embedding models align with the ratings of the human annotators using statistical measures. Then, we apply our emoji embedding models to a sentiment analysis task to demonstrate the utility of them in a real-world NLP application. Our models were able to correctly predict the sentiment class of tweets laden with emoji from a benchmark dataset [Novak et al. 2015] with an accuracy of 63.6% (7.73% improvement), outperforming recent results on the same dataset [Barbieri et al. 2016; Eisner et al. 2016].

1.1.4 Emoji Sense Disambiguation

Emoji sense disambiguation is the ability to identify the meaning of an emoji in the context of a message in a computational manner. Previous research has identified the importance of the problem [Wijeratne et al. 2016], however, have not solved it. To solve the emoji sense disambiguation problem, there has to be an emoji sense inventory that a computer program could use to extract

emoji meanings. We hypothesize that EmojiNet can be used to solve emoji sense disambiguation problem.

To discuss how EmojiNet can be used to solve emoji sense disambiguation problem, we provide an illustration of disambiguation of the sense of the 🙏 emoji as it is used in two example tweets. We choose this emoji since it is reported as one of the most misunderstood emoji on social media¹⁵. The tweets we consider are:

T_5 : Pray for my family 🙏 God gained an angel today.

T_6 : Hard to win, but we did it man 🙏 Lets celebrate!

EmojiNet lists **high five(noun)** and **pray(verb)** as valid senses for the above emoji. For **high five(noun)**, EmojiNet lists three definitions and for **pray(verb)**, it lists two definitions. We take all the words that appear in their corresponding definitions as possible context words that can appear when the corresponding sense is being used in a sentence (tweet in this case). For each sense, EmojiNet extracts the following sets of words:

pray(verb) : {*worship, thanksgiving, saint, pray, higher, god, confession*}

highfive(noun) : {*palm, high, hand, slide, celebrate, raise, person, five*}

To calculate the sense of the 🙏 emoji in each tweet, we calculate the overlap between the words which appear in the tweet with words appearing with each emoji sense listed above. This method is called the Simplified Lesk Algorithm [Vasilescu et al. 2004]. The sense with the highest word overlap is assigned to the emoji at the end of a successful run of the algorithm. We can see that 🙏 emoji in T_5 will be assigned **pray(verb)** based on the overlap of words {god, pray} with words retrieved from the sense definition of **pray(verb)** and the same emoji in T_6 will be assigned **high five(noun)** based on the overlap of word {celebrate} with words retrieved from the sense definition of **high five(noun)**. In the above example, we have only shown the minimal set of words that one could extract from EmojiNet. Since we link EmojiNet senses with their corresponding BabelNet

¹⁵<http://www.goodhousekeeping.com/life/g3601/surprising-emoji-meanings/>

senses using BabelNet sense IDs, one could easily utilize other resources available in BabelNet such as related WordNet¹⁶ senses, VerbNet¹⁷ senses, Wikipedia, etc. to collect an improved set of context words for emoji sense disambiguation tasks. Emoji sense disambiguation has applications in other downstream tasks such as sentiment and emotion analysis.

1.2 Dissertation Organization

The rest of the dissertation is organized as follows. Chapter 2 presents an overview of the past research conducted on emoji understanding. First, it tries to position emoji in the semiotics (i.e., the general body of research on how meanings are assigned to symbols) space. Then, it discusses the past research related to emoji understanding, emoji similarity, emoji usage and applications, and emoji sense disambiguation. It also frames how the work presented in this dissertation differs from and furthers existing research.

Chapter 3 presents the creation of EmojiNet framework in detail. It discusses how an emoji is represented using a nine-tuple notation, the data extraction process from online emoji websites, data validation process, and the data integration process used to create EmojiNet. It also presents the evaluation of data cleaning and integration processes. It further discusses EmojiNet Application Programming Interface (API) along with supported web services calls and links to download EmojiNet datasets.

Chapter 4 presents the work conducted on emoji similarity calculation. It first discusses how the emoji embeddings are learned by incorporating emoji meanings from EmojiNet. Then it discusses the creation of EmoSim508 dataset, presents the evaluation results for emoji embeddings, and shows that combining distributional semantics with existing knowledge can improve emoji embeddings.

Chapter 5 presents the work on emoji sense disambiguation. It highlights the importance of disambiguating the emoji meaning in social media posts and presents methodologies to disambiguate

¹⁶<https://wordnet.princeton.edu/>

¹⁷<https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

emoji meanings in text. It also discusses two key takeaways learned from sense disambiguation experiment, namely, (i) longer context lengths improve the disambiguation accuracy, and (ii) tools designed for well-formed text processing will not work well when used for ill-formatted text processing.

Chapter 6 concludes the dissertation by summarizing the key contributions and insights. It also discusses future research in the area of building emoji sense inventories and emoji understanding tasks. It also shows how the work presented in this dissertation can be further extended.

2

Background and Related Work

2.1 Overview

This chapter reviews research related to building emoji sense inventories, emoji similarity calculation, and emoji sense disambiguation. First, it discusses emoticons and emoji in brief. Then, it tries to position emoji in the semiotics space and discusses the common emoji usages. Finally, it frames how the work presented in this dissertation differs from other related works discussed.

2.2 Emoticons, Emoji, and Other Pictographs

We first look at emoticons and how they differ from emoji. Emoticons or smiley characters are pictorial representations of facial expressions, created by using punctuation marks, such as :-)) and :-(. Even though the general concept of ‘smiley’ characters was introduced in the mid-1960s with their pictorial representations [Danesi 2016], the emoticons that use punctual marks that are largely popular today were introduced by Scott Fahlman in 1982¹. Since then, emoticons have been widely used for expressing emotion in online communications [Wang et al. 2012; Wang 2015]. However, recent studies show that emoji are slowly taking over emoticons in online communications [Dimson

¹<https://en.wikipedia.org/wiki/Emoticon>

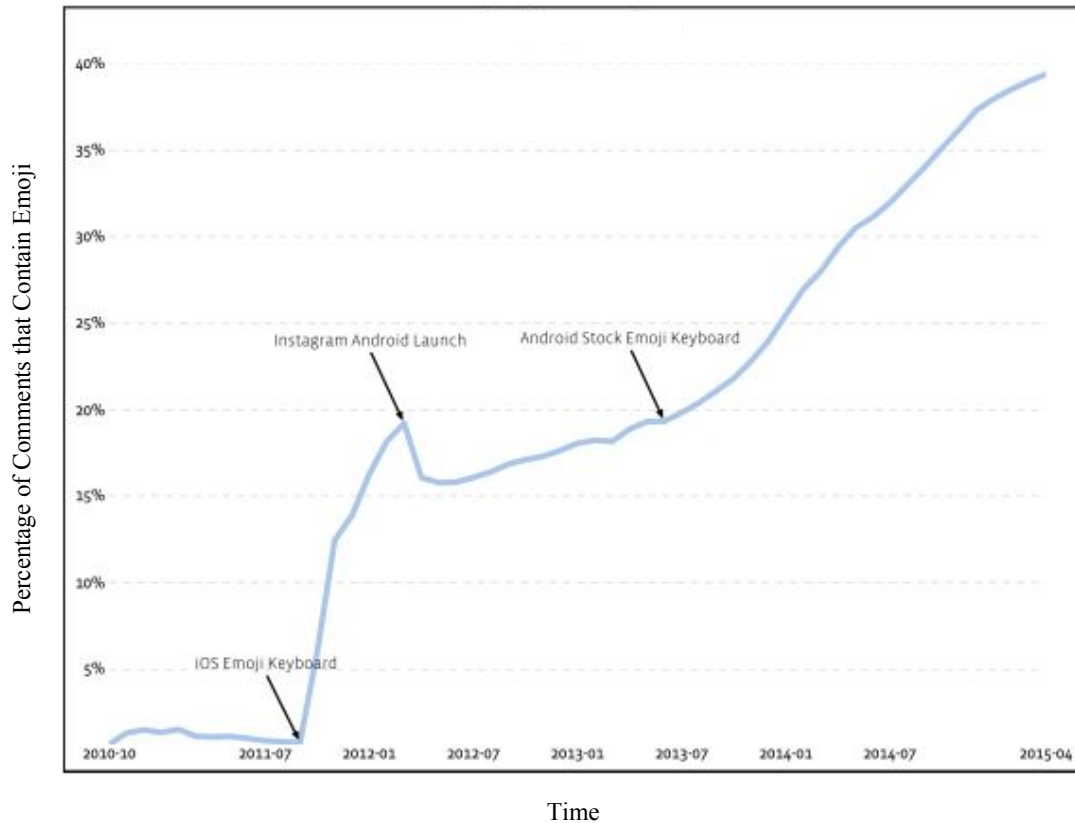


Figure 2.1: Emoji Usage Growth on Instagram. Image Source – <http://bit.ly/2QVaKTS>.

2015; Pavalanathan and Eisenstein 2015; 2016].

Emoji were first introduced by Shigetaka Kurita in the late 1990s for a Japanese telecommunication organization named NTT DoCoMo². He proposed a set of 176 emoji characters, however, they were not widely adopted by other telecommunication providers around the world, except for the Japanese market. After the Unicode Consortium standardized emoji characters in 2010, telecommunication, mobile, and web platform providers quickly adopted the new Unicode emoji code points into their platforms. For example, Apple released an emoji keyboard in 2011 for Apple devices and Android did the same in 2013 for Android devices. These emoji keyboards let the users to easily incorporate emoji into their social media messages. For example, a study by Instagram shows that

²<https://www.cnn.com/style/article/emoji-shigetaka-kurita-standards-manual>

more than 40% of Instagram photo comments in the year 2015 contained at least one emoji [Dimson 2015]. This emoji growth is also shown in Figure 2.1.

Apart from emoji, there are many other pictographic representations proposed and used by different mobile and web platform providers. For example, mobile application providers such as Bitmoji³ and video platforms such as Twitch⁴ have come up with their own sets of emoji-like pictographic characters. Similarly, Facebook has come up with their own versions of graphicons (also known as stickers). Researchers have studied how these pictographs have been used in their corresponding online social media platforms [Barbieri et al. 2017; Jiang et al. 2017; Herring and Dainas 2017]. However, these pictographs are not standardized in such a way that they can be used across multiple web and social media platforms. Hence, it is not possible to develop methods which utilize traditional NLP methods that generalize across multiple platforms. Thus, the focus of this dissertation is limited to the emoji characters that are standardized by The Unicode Consortium.

2.3 Semiotics, Writing Systems, and Emoji

Semiotics is defined as “the study of signs” [Chandler 2007]. The first reference to semiotics as a branch of philosophy traces back to John Locke’s (1632–1704) article “An Essay Concerning Human Understanding” published in 1690 [Locke 1841]. Following Locke, many philosophers and linguists studied how symbols are given meanings. Among them, the Swiss linguist Ferdinand de Saussure (1857–1913) and the American philosopher Charles Sanders Peirce (1839–1914), who are considered as the co-founders of semiotics [Chandler 2007], contributed to the two widely accepted definitions for semiotics.

Coming from a linguistic background, Saussure defined that a sign composed of two parts; a **signifier** and a **signified**. Signifier describes the form that the sign takes (e.g., a word such as ‘fire’) and the signified describes the concept that the signifier refers to. In Saussure’s model of a sign,

³<https://www.bitmoji.com/>

⁴<https://www.twitch.tv/>

the presence of signifier and signified is mandatory and the signifier can stand for different signifieds (e.g., the word ‘fire’ can take different meanings in different contexts). Similarly, many signifiers can stand for a given form of another signifier (e.g., the sign forms ‘combustion’ and ‘burning’ can also be signifiers for the sign form ‘fire’).

Peirce proposed a triadic model to define a sign. The three parts of Peirce’s model are:

The representamen :the ‘sign vehicle’ or what is represented by the sign (similar to signifier in Saussure’s model but broader).

An interpretant :the meaning of the sign.




An object :what is being referred to by the sign (similar to signified in Saussure’s model).

Peirce further stated that the relationship between **representamen** and **object** can take three forms. They are:

Symbol/Symbolic :the sign represents the object through an arbitrary/conventional relation, thus, the relation should be agreed upon or learned.

Icon/Iconic :the sign resembles or imitates the object.

Index/Indexical :the signifier is directly connected through a form of association.

Researchers have tried to explain emoji usage through the semiotics models discussed above. Specifically, Bai [Bai 2018] tries to explain the assignment of meanings to emoji symbols via symbolic, iconic, and indexical relationships proposed by Peirce. In [Bai 2018], Bai provides examples for such usages. He argues that the ‘guardsman emoji’  gets the meaning ‘guardsman’ simply because the emoji resembles a guardsman, which follows the iconic form. The heart emoji  is commonly associated with the meaning ‘love’ as heart shape is already associated with that meaning [Bai 2018], which follows the symbolic form. He lists the ‘peach emoji’  as an example for indexical form as peach emoji refers to multiple objects.

Researchers have also studied how people use emoji and reported that the emoji are used mainly

to express sentiment [Danesi 2016; Cramer et al. 2016], to replace words [Danesi 2016; Na’aman et al. 2017; Donato and Paggio 2017], and to emphasize words in a sentence [Danesi 2016; Donato and Paggio 2017]. These emoji usage patterns show resemblance to different writing systems. For example, Marcel Danesi argues that emoji show pictographic (i.e., representation of objects) and logographic (i.e., word replacement) functions of the language systems [Danesi 2016]. In [Danesi 2016], Danesi discusses the cast of language systems from early pictographic languages to logographic and alphabetic languages. He argues that the use of emoji with alphabetic languages could lead to a new paradigm shift in the language where pictographic-logographic writing coexist with alphabetic writing. We have already started to experience this new form of writing style in online social media platforms. However, researchers admit that emoji would not contribute to a whole new language [Evans 2017] of its own⁵, even though they have shown that emoji express certain features of languages. For example, McCulloch *et al.* studied the repetitive emoji patterns to draw insights into emoji grammar [McCulloch and Gawne 2018]. After studying the sequences of most common two, three, and four emoji strings, they reported that emoji-based communications do not show the characteristics of grammars with hierarchical structures. They stated that the repetitive emoji resembles beat gestures, which is a well-established type of co-speech gesture characterized by its high level of repetition.

2.4 Emoji Understanding and Building Emoji Sense Dictionaries

Although emoji was introduced two decades ago, research on this communication form remains limited [Miller et al. 2016]. Emoji became popular when mobile service vendors adopted emoji onto mobile platforms [SwiftKey 2015]. Early research into emoji focused on understanding the role of emoji in computer-mediated communication. Cramer *et al.* studied the sender-intended functionality

⁵<http://www.bbc.com/future/story/20151012-will-emoji-become-a-new-language>

of emoji using a group of 228 individuals who used emoji in text messages [Cramer et al. 2016]. They reported on functional differences in emoji use and showed that the social and linguistic functions of emoji are complex. They argued that the functions of emoji go beyond expressing emotions in online communications. They reported that people use emoji to provide additional emotional and situational information, as a means of tone modification (e.g., using a face emoji at the end of the text to bring down the tone of the message), emphasize certain words in a message (e.g., Happy Birthday 🎂), and for text/word replacement (e.g., I ❤️ you). They also highlighted the importance of building natural language processing systems to understand emoji, which is a primary focus of this dissertation. Hu *et al.* also studied how emoji are interpreted by senders and receivers in online communications [Hu et al. 2017]. They reported that senders commonly use emoji to expressing the sentiment, strengthening expression, and adjusting the tone of messages. These results align with Cramer *et al.*'s observations [Cramer et al. 2016].

Kelly *et al.* studied how people in close relationships use emoji. They reported that people who are in close relationships use emoji as a way of maintaining conversational connections in a playful manner [Kelly and Watts 2015]. Similar to the study by Cramer *et al.* [Cramer et al. 2016], Kelly *et al.* also reported that the functions of emoji in online communication go beyond simple emotion expression. They reported that people in close relationships often create a shared and secret bond within a particular relationship by using emoji [Kelly and Watts 2015]. Pavalanathan *et al.* studied how emoji compete with ASCII-based non-standard orthographies, including emoticons, when it comes to communicating paralinguistic content on social media [Pavalanathan and Eisenstein 2016]. They reported that Twitter users prefer emoji over emoticons, and users who adopt emoji tend to use standard English words at an increased rate after emoji adoption. Their experiments revealed that emoticons with horizontal orientation and winking eyes are relatively less used after the introduction of emoji. They also argued that if people continue to use emoji to replace emoticons and other non-standard orthographies, emoji could, in fact, solve the problems with using non-standard orthographies in online communications.

Past work on understanding emoji meanings by Miller *et al.* focused on how the sentiment and semantics of emoji differ when the same emoji is displayed on multiple platforms as vendors can design their own emoji image to display [Miller et al. 2016]. In their experiments, they reported that even after seeing the same emoji rendering, the participants were confused about the correct sentiment of the emoji 25% of the time. They also reported that the disagreement in the sentiment of the emoji only increase when considering the renderings across different platforms. In their follow-up work, they studied whether the text surrounded by emoji can be used to determine the sentiment of emoji [Miller et al. 2017]. They reported that miscommunication can still exist even when emoji are interpreted in textual contexts. They also noted that emoji sense disambiguation, which is a problem studied in this dissertation is a challenging, hard-to-solve task [Miller et al. 2017].

Tigwell *et al.* also studied how emoji misunderstanding can happen due to platform-specific emoji designs [Tigwell and Flatla 2016]. They surveyed people about their use of emoji to investigate the variation in their interpretation of emoji. Specifically, they looked at people’s interpretations of emoji available in Android and iOS platforms, which are the two most popular mobile platforms. They argued that systems should be built in such a way that they can transfer the sender’s intention along with emoji to solve emoji-related miscommunications. Researchers have also explored whether different genders interpret emoji differently and reported that there are no significant differences in the way males and females interpret emoji [Herring and Dainas 2018].

Several researchers have worked on building semantic models that can be used to understand and interpret emoji in computer applications. For example, Barbieri *et al.* studied emoji meanings using word embeddings [Mikolov et al. 2013] learned over a tweet corpus and used the learned word embeddings to calculate the functional and topical similarity between emoji [Barbieri et al. 2016]. They showed that the emoji embeddings they learned over words surrounding the emoji in text messages can be used to understand the relationships among emoji. They used t-Distributed Stochastic Neighbor Embedding (t-SNE) method [Maaten and Hinton 2008] to visualize the learned embedding models and showed that the emoji embedding models can be used to obtain meaningful

emoji clusters that can be used to calculate emoji similarity. Eisner *et al.* used emoji descriptions available on the Unicode Consortium Website to learn emoji meanings [Eisner et al. 2016] and showed that their emoji representation model could outperform Barbieri *et al.*'s model in a sentiment analysis task. Eisner *et al.*'s approach was able to combine the emoji meaning knowledge available in Unicode Consortium Website with word embedding models. They also showed that emoji embeddings can be used to improve real-world natural language processing tasks such as sentiment analysis. Pohl *et al.* [Pohl et al. 2017] also developed a similar emoji embedding model to Barbieri *et al.* [Barbieri et al. 2016] and they argued that emoji embeddings can be used to calculate emoji similarity, which can then be used to optimize emoji keyboard designs for hand-held devices. Ai *et al.* used emoji embeddings to study the correlation between emoji semantics and emoji usage [Ai et al. 2017]. They showed that emoji popularity is affected by several factors including the structural properties of emoji, how complementary they are to the words and the sentiment of the context that they are being used. Illendula *et al.* proposed an emoji embedding model based on emoji co-occurrence. They generated an emoji co-occurrence network based on a large corpus of tweets and showed that emoji embeddings learned over emoji co-occurrence graphs could improve the performance of downstream natural language processing applications such as sentiment analysis [Illendula and Yedulla 2018].

Past research has shown that emoji can be used as features for sentiment and emotion analysis experiments [Wijeratne et al. 2017]. For example, Novac *et al.* developed an emoji sentiment ranking model based on the sentiment associated with emoji in the messages used in online communications [Novak et al. 2015]. After analyzing the sentiment of 1.6 million tweets in 13 European languages, they reported that the sentiment of the popular face emoji are generally positive. They also reported that the sentiment distribution of the tweets with and without emoji can be significantly different, suggesting that emoji are extensively used to express the sentiment in online communications. They further reported that emoji tend to occur at the end of the text messages and their sentiment polarity increases with the distance. Others have used features extracted from emoji usage in various computational problems including emotion analysis [Wang et al. 2012; Wood

and Ruder 2016], understanding communication in disaster events [Santhanam et al. 2018], and Twitter profile classification [Balasuriya et al. 2016; Wijeratne et al. 2016].

Work on building resources that enable the natural language interpretation of emoji is at a very early stage. Several web resources list emoji senses either as keywords or sense labels, which is defined as a `word(PoS tag)` pair such as `laugh(noun)`. Sense labels can be helpful for developing emoji sense inventories. For example, The Unicode Consortium⁶ provides lists of keywords that could act as the intended meanings for emoji. The Emoji Dictionary lists sense labels for emoji meanings that are collected via crowdsourcing. However, none of these web resources can serve as machine-readable sense inventories due to the limitations in their system designs, including not providing enough training examples for a computer program to understand how an emoji should be used in a message context [Wijeratne et al. 2016; 2017a]. Therefore, simply scraping those websites to extract emoji sense labels alone cannot help to build emoji sense inventories. The sense labels need to be linked with machine processable dictionaries such as BabelNet [Navigli and Ponzetto 2010] to extract message contexts for them. The Emoji Dictionary contains more valuable emoji meanings compared to what the Unicode Consortium website has to offer, but The Emoji Dictionary does not list Unicode code points of emoji, which makes it difficult to be directly consumed by a computer program. Thus, this dissertation focuses on the construction of a machine-readable emoji sense inventory and using it to solve emoji understanding tasks.

2.5 Emoji Similarity

Emoji similarity has received little attention apart from three attempts by Barbieri *et al.* [Barbieri et al. 2016], Eisner *et al.* [Eisner et al. 2016] and Pohl *et al.* [Pohl et al. 2017]. Barbieri *et al.* [Barbieri et al. 2016] collected a sample of 10 million tweets originated from the USA and trained an emoji embedding model using tweets as the input. Then, using 50 manually-generated emoji pairs

⁶<https://goo.gl/1o3z1E>

annotated by humans for emoji similarity and relatedness, they evaluated how well the learned emoji embeddings align with the human annotations. They reported that the learned emoji embeddings align more closely with the relatedness judgment scores of human annotators than the similarity judgment scores. Eisner *et al.* [Eisner et al. 2016] used a word embedding model learned over the Google News corpus⁷, applied it to emoji names and keywords extracted from the Unicode Consortium website, and learned an emoji embedding model which they called `emoji2vec`. Using t-SNE for data visualization [Maaten and Hinton 2008], Eisner *et al.* showed that the high dimensional emoji embeddings learned by `emoji2vec` could group emoji into clusters based on their similarity. They also showed that their emoji embedding model could outperform Barbieri *et al.*'s model in a sentiment analysis task. Pohl *et al.* [Pohl et al. 2017] studied the emoji similarity problem using two methods; one based on the emoji keywords extracted from the Unicode Consortium website and the other based on emoji embeddings learned from a Twitter message corpus. They used the Jaccard Coefficient⁸ on the emoji keywords extracted from the Unicode Consortium to find the similarity of two emoji. They evaluated their approach using 90 manually-generated emoji pairs and discussed how emoji similarity can be used to optimize the design of emoji keyboards.

The work presented in this dissertation differs from the related works discussed above in many ways. Barbieri *et al.* [Barbieri et al. 2016] use the distributional semantics [Harris 1954] of words learned over a Twitter corpus where they seek an understanding of emoji meanings from how emoji are used in a large collection of tweets. In contrast, we learn emoji embeddings based on emoji meanings extracted from EmojiNet. We learn the distributional semantics of the words in emoji definitions using word embeddings learned over two large text corpora and use the learned word embeddings to model the emoji meanings extracted from EmojiNet. Hence, we combine emoji meanings extracted from knowledgebases (i.e., EmojiNet) with distributional semantics of those words in emoji definitions. Pohl *et al.* [Pohl et al. 2017] learn emoji embedding models in the same

⁷<https://goo.gl/QaxjVC>

⁸<https://goo.gl/RKkRzF>

way as Barbieri *et al.* and use the Jaccard Coefficient⁹ on emoji keywords extracted from the Unicode Consortium to measure similarity. This is similar to our earlier work on emoji similarity which is discussed in [Wijeratne et al. 2017a], and we further extend that work in this dissertation. Eisner *et al.*'s [Eisner et al. 2016] presented an embedding model built on short emoji names and keywords listed on the Unicode Consortium website, which is approximately 4 to 5 words long on average as reported by Pohl *et al.* in [Pohl et al. 2017]. Since prior research suggests that the emoji embedding models can be improved by incorporating more words by using longer emoji definitions [Eisner et al. 2016; Pohl et al. 2017], we introduce embeddings based on three different types of long-form definitions of an emoji. We show that our embedding models outperform the other emoji embedding models in a downstream sentiment analysis task [Wijeratne et al. 2017b].

2.6 Emoji Sense Disambiguation

Previous research has identified the importance of emoji sense disambiguation problem [Miller et al. 2016; Miller et al. 2017], however, have not solved it. Specifically, past research has looked at creating datasets that can be used to identify the linguistic roles of emoji use in online communications. For example, Donato *et al.* looked at tweets with emoji and annotated them based on the linguistic roles of emoji [Donato and Paggio 2017]. They identified three linguistic roles of emoji, namely, (i) redundant, (ii) non-redundant, and (iii) non-redundant with PoS. The redundant category consists of tweets where the emoji were used to emphasize words in the tweet (e.g., We'd love to have birthday cake! 🎂). The non-redundant category consists of tweets where emoji refer to information that are not present in the tweet (e.g., I wish you were here ✈️). Non-redundant with PoS category consists of tweets where emoji were used to replace the words in the tweet (e.g., We love eating 🍕). Even though Donato *et al.* created the annotated datasets, they didn't work on building models that can identify those emoji functions in text. Na'aman *et al.* also looked at the functions of emoji in

⁹https://en.wikipedia.org/wiki/Jaccard_index

textual communications [Na’aman et al. 2017] and identified emoji functions that are similar to the ones reported by Donato *et al.* [Donato and Paggio 2017].

Miller *et al.* studied whether the text surrounded by emoji can be used to determine the sentiment of emoji [Miller et al. 2017]. They reported that miscommunication can still exist even when emoji are interpreted in textual contexts and reported that emoji sense disambiguation is a challenging, hard-to-solve problem [Miller et al. 2017]. We discussed the challenges in solving emoji sense disambiguation problem in a supervised setting in Section 1.1.2 and showed how emoji meaning dictionaries can be used to overcome the challenges with supervised methods. There are several online web resources available on the web that can act as emoji dictionaries. The Emoji Dictionary¹⁰ is a promising Web resource that could be utilized by humans for emoji sense disambiguation. It is a crowdsourced emoji dictionary that provides emoji definitions with user defined sense labels, which are `word(PoS tag)` pairs such as `laugh(noun)`. However, it cannot be utilized by a machine for several reasons. First, it does not list the Unicode or shortcode names for emoji, which are common ways to programmatically identify emoji characters in text. Secondly, it does not list sense definitions and example sentences along with different sense labels for emoji. Typically, when using machine readable dictionaries, machines use such sense definitions and example sentences to generate contextually relevant words for each sense in the dictionary. Thirdly, the reliability of the sense labels is unclear as no validation of the sense labels submitted by the crowd is performed. With EmojiNet, we address these limitations by linking The Emoji Dictionary with other rich emoji resources found on the Web. This allows sense labels to be linked with their Unicode and shortcode name representations and discards human-entered sense labels for emoji that are not agreed upon by the resources. EmojiNet also links sense labels with BabelNet to provide definitions and example usages for different senses of an emoji [Wijeratne et al. 2016; 2017a].

¹⁰<http://emojidictionary.emojifoundation.com/home.php?learn>

2.7 Summary

This chapter presented an overview of the past research conducted on emoji understanding. First, it positioned emoji in the semiotics space and discussed the past research related to emoji understanding. Then it framed how the work presented in this dissertation differs from and furthers existing research. It also discussed the related work on emoji research, which is essential to understand the concepts discussed in the next chapters of this dissertation.

3


EmojiNet: Building a Machine-Readable Emoji Sense Inventory

3.1 Overview

This chapter present EmojiNet, the first machine-readable emoji sense inventory that maps emoji to their set of possible meanings or *senses*. It discusses how an emoji is modeled in EmojiNet, the processes involved in creating it, and reports on the evaluation matrices used. It also discusses an application use-case of EmojiNet where emoji meanings are used to determine the similarity of emoji pairs¹.

¹Content presented in this chapter were previously published in [Wijeratne et al. 2017a]

Table 3.1: Nonuple Representation of an Emoji.

Nonuple Element	Description
Unicode u_i	U+1F64C
Emoji Name n_i	Raising Hands
Short Code c_i	:raised_hands:
Definition d_i	Two hands raised in the air, celebrating success or an event.
Keywords K_i	celebration, hand, hooray, raised
Images I_i	
Related Emoji R_i	Confetti Ball, Clapping Hands Sign
Emoji Category H_i	Gesture symbols
Senses S_i	Sense Label: celebration(Noun) Def: A joyful occasion for special festivities to mark a happy event.

3.2 Emoji Modeling and Dataset Creation

EmojiNet exposes a dataset of emoji. Each emoji is represented as a nonuple representing its sense and other metadata. Let E be the set of all emoji in EmojiNet. For each emoji $e_i \in E$, EmojiNet records the nonuple $e_i = (u_i, n_i, c_i, d_i, K_i, I_i, R_i, H_i, S_i)$, where u_i is the Unicode representation of e_i , n_i is the name of e_i , c_i is the short code of e_i , d_i is a description of e_i , K_i is the set of keywords that describe intended meanings attached to e_i , I_i is the set of images that are used in different rendering platforms, R_i is the set of related emoji extracted for e_i , H_i is the set of categories that e_i belongs to, and S_i is the set of different senses in which e_i can be used within a sentence.

An example of nonuple notation is shown in Table 3.1. Each element in the nonuple provides essential information on emoji and for emoji sense disambiguation. EmojiNet uses unicode u_i , name n_i , and short code name c_i of an emoji $e_i \in E$ to uniquely identify e_i , and hence, to search EmojiNet. d_i is a description of what is modeled in the emoji. It can sometimes help to understand the intended

use of an emoji too. K_i is also helpful to understand the intended uses of an emoji. I_i helps to understand the rendering differences in each emoji based on different platforms. R_i and H_i could be useful to understand how emoji are related; thus, will be useful in tasks such as calculating emoji similarity and emoji sense disambiguation. Finally, S_i holds all senses for e_i , including their POS tags and sense definitions and links them with BabelNet, which makes EmojiNet a machine-readable emoji sense inventory.

3.3 Open Resources Used in EmojiNet

A number of open resources, with appropriate permission from the dataset owners, are used to construct the nonuple of an emoji. This section introduces those resources and the information extracted from each of them.

3.3.1 The Unicode Consortium Emoji List

Unicode is an industry standard for the encoding, representation, and handling of text in computers which enables people around the world to use them in any language². The Unicode Consortium also maintains a complete list of the standardized Unicodes for each emoji³ along with other information on them such as manually curated keywords and images of emoji. Let the set of all emoji available in the Unicode emoji list be E_U . For each emoji $e_u \in E_U$, the Unicode character u_u of e_u , the name n_u of e_u , the set of all manually assigned keywords K_{e_u} that describe the intended functionality of e_u , the set of all images I_{e_u} associated with e_u that are used to display e_u on different platforms, and the set of categories H_{e_u} which are all the categories that e_u belongs to, are extracted from the Unicode Consortium website for inclusion in EmojiNet.

²<http://www.unicode.org/>

³<https://goo.gl/1o3z1E>

3.3.2 Emojipedia

Emojipedia⁴ is a human-created emoji reference website that organizes emoji into a pre-defined sets of categories while also providing useful information about them. Specifically, for each emoji, Emojipedia lists the Unicode representation of the emoji, its short code, its variations over rendering platforms, and its relationships with other emoji. Let the set of all emoji available in Emojipedia be E_E . For each emoji $e_e \in E_E$, Emojinet extracts the Unicode representation u_e , short code c_e , emoji definition d_e , and the set of related emoji R_{e_e} of e_e from Emojipedia.

3.3.3 The Emoji Dictionary

The Emoji Dictionary⁵ is the first crowdsourced emoji reference website that provides emoji definitions with their sense labels based on how they are used in sentences. It organizes the different meanings of an emoji under three part-of-speech tags, namely, nouns, verbs, and adjectives. It also lists an image of the emoji and its definition with example usages spanning across multiples senses with multiple part-of-speech tags. Let the set of all emoji available in The Emoji Dictionary be E_D . For each emoji $e_d \in E_D$, Emojinet extracts its image $i_{e_d} \in I_D$, where I_D is the set of all images of all emoji in E_D and the set of crowd-generated sense labels S_{e_d} from Emoji Dictionary.

3.3.4 BabelNet

BabelNet⁶ is the most comprehensive multilingual machine-readable semantic network available to date [Navigli and Ponzetto 2010] and it has been shown useful in many research areas, including word sense disambiguation [Navigli 2009; Navigli and Ponzetto 2010], semantic similarity [Camacho-Collados et al. 2015], and sense clustering [Camacho-Collados et al. 2015]. It is a dictionary with a lexicographic and encyclopedic coverage of words within a semantic network that connects concepts in Wikipedia to the corresponding words in the BabelNet dictionary. Sense definitions from BabelNet

⁴<http://emojipedia.org/>

⁵<https://goo.gl/9gDVkE>

⁶<https://babelnet.org/>

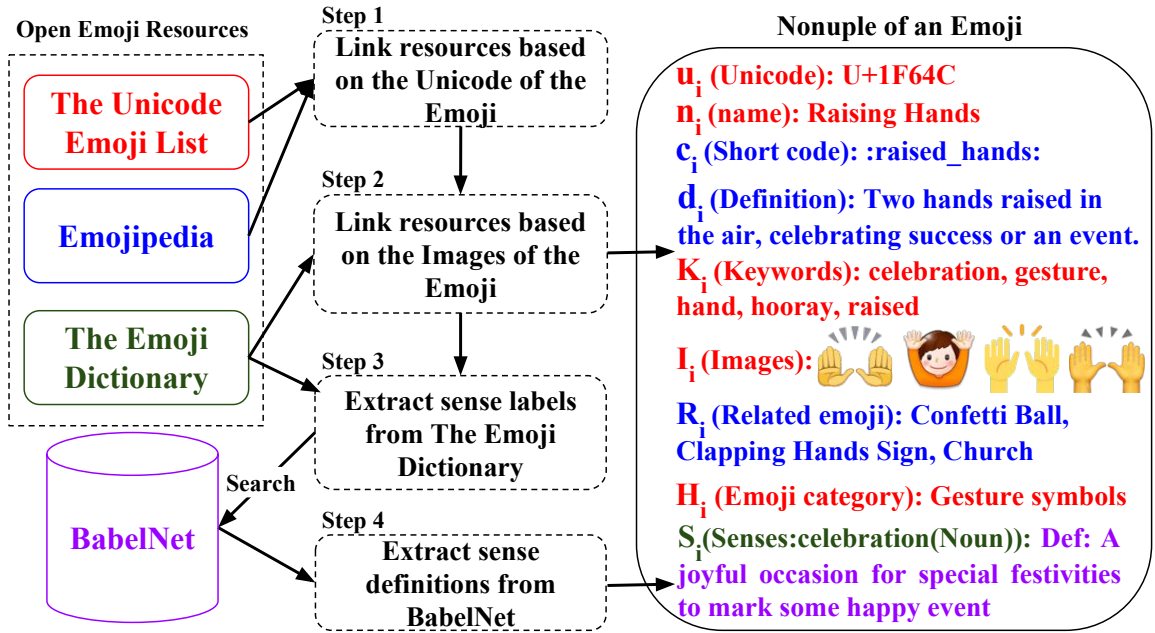


Figure 3.1: Construction of Emoji Representation in EmojiNet.

are included in EmojiNet. For the set of all sense labels S_{e_d} in each $e_d \in E_D$, EmojiNet extracts the sense definitions and examples for each sense label $s_{e_d} \in S_{e_d}$ from BabelNet.

3.4 Resource Integration

Figure 3.1 gives a high level four step overview of how the open resources are utilized to create an emoji representation. This section elaborates on each of these steps.

3.4.1 Linking Resources based on the Unicode Code Points

First, EmojiNet extracts all emoji characters that are currently supported by the Unicode Consortium and the information it stores for each one of them, such as emoji names, keywords and images. Then, for each emoji extracted from the Unicode website, EmojiNet extracts additional information such as the emoji short code, emoji description and related emoji from Emojipedia website. EmojiNet merges all information extracted from the two websites based on the Unicode representation

of emoji and stores them under each emoji $e_i \in E$ using the nonuple notation described earlier.

3.4.2 Linking Resources based on the Images

The Emoji Dictionary does not store the Unicode character representations of emoji, hence integrating it with the emoji data extracted from the Unicode Consortium and Emojipedia websites is done based on matching the images of the emoji available in the three resources. For this purpose, we extracted 18,615 images representing all emoji extracted from the Unicode Consortium website and created an index, which we refer to as our example set, I_x . We also downloaded images of all emoji listed on The Emoji Dictionary website, which resulted in a total of 1,074 images, from which we created our test image dataset, I_t . We implemented a nearest neighborhood-based image matching algorithm based on [Santos 2010] that matches each image in I_t with the images in I_x . This algorithm has shown to perform well when aligning images with few colors and objects, which is the case with emoji. Since images are of different resolutions, we first normalized them into a 300x300px space and then divided them along a lattice of 25 non-overlapping regions of size 25x25px. We then calculated the average color intensity of each region by averaging its R , G and B pixel color values. To calculate the dissimilarity between two images, we summed the L_2 distance of the average color intensities of the corresponding regions. We selected L_2 distance as it prefers many medium disagreements to one large disagreement as in L_1 distance. The final accumulated value that we received for a pair of images will be a measure of the dissimilarity of the two images. For each image in I_t , the least dissimilar image from I_x is chosen and the corresponding emoji nonuple information is merged.

3.4.3 Extracting Sense Labels

The Emoji Dictionary lists sense labels for each emoji which were obtained through its crowdsourced data collection platform, while the Unicode Consortium lists intended meanings for each emoji as keywords, but without any part-of-speech tags. The two resources thus carry complementary

information about emoji meanings necessary to create a sense label. EmojiNet follows the procedure illustrated in Figure 3.2 to extract emoji sense labels using the two resources. For each emoji, EmojiNet extracts all the emoji sense labels listed in The Emoji Dictionary. This resulted in a total of 31,944 sense labels. For each of the 6,057 keywords for emoji listed in the Unicode Consortium website, it then generates three sense labels using the three part-of-speech tags; noun, verb, and adjective. That means, for a keyword listed in the Unicode Consortium website such as **face**, EmojiNet generates the three sense labels **face(N)**, **face(V)**, and **face(A)** as shown in Figure 3.2. We selected only three part-of-speech tags as they were the only part-of-speech tags supported by The Emoji Dictionary and other emoji sense inventories [Wijeratne et al. 2016].

Following the above step, a total of 18,171 sense labels for the 6,057 keywords are created. Next, EmojiNet combines the sense labels from the two resources into a pool of sense labels, totaling 50,115 sense labels in the pool. However, not all senses in the pool of sense labels are valid. For example, the sense label **face(A)** is invalid as the word **face** cannot be used as an adjective in the English language. To filter out invalid sense labels from the sense label pool, EmojiNet validates each sense label in the pool against the valid sense labels in BabelNet sense inventory. During this validation process, a total of 21,779 sense labels were discarded from the sense label pool where 10,848 of them were extracted from The Emoji Dictionary and 10,931 of them were generated from the Unicode Consortium keywords. The above filtering step leaves 28,336 valid sense labels in the sense label pool. We also noticed that there are a lot of sense labels in the pool that do not represent valid meanings for certain emoji. For example, for the emoji 🐷, **pig(N)**, **rainbow(N)**, and **face(V)** are listed among many other invalid meanings. Even though these are valid English sense labels, they are not valid meanings for the 😄 emoji, thus we remove such instances. Most of such invalid sense labels were extracted from The Emoji Dictionary, and due to non availability of input validation methods in The Emoji Dictionary website, those being ended up adding to Emoji Dictionary’s sense inventory. With the help of two human annotators, we were able to remove a total of 15,432 such sense labels. The remaining 12,904 sense labels are ready to be assigned their sense definitions using

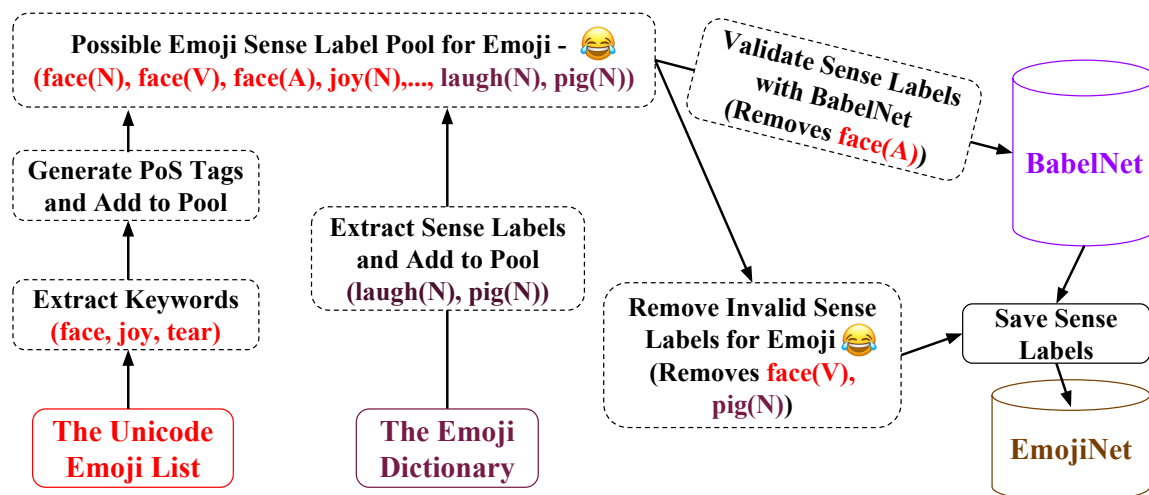


Figure 3.2: Using The Unicode Consortium and The Emoji Dictionary for Sense Label Filtering.

BabelNet, as described next.

3.4.4 Extracting and Linking with BabelNet Senses

For a given sense label, there could be multiple sense definitions available in BabelNet. For example, the current version of BabelNet lists 6 different sense definitions for the sense label `laugh(noun)`. Thus, to select the most appropriate sense definition out of the multiple BabelNet sense definitions and link them with the sense labels extracted in the earlier step, a Word Sense Disambiguation (WSD) task needs to be performed. To conduct this WSD task, we use the Manually Annotated Sub-Corpus (MASC)⁷ with a most frequent sense (MFS) baseline. We choose the MASC corpus because it is a balanced dataset that represents different text categories such as tweets, blogs, emails, etc. and Moro *et al.* have already annotated it using the BabelNet senses [Moro et al. 2014]. A MFS baseline outputs the MFS calculated for each word with respect to a sense-annotated text corpus. Then the baseline assigns the MFS of a word as its correct word sense. We use a MFS-based WSD baseline due to the fact that MFS is a very strong, hard-to-beat baseline model for WSD tasks [Basile et al. 2014].

⁷<https://goo.gl/0eLc2F>

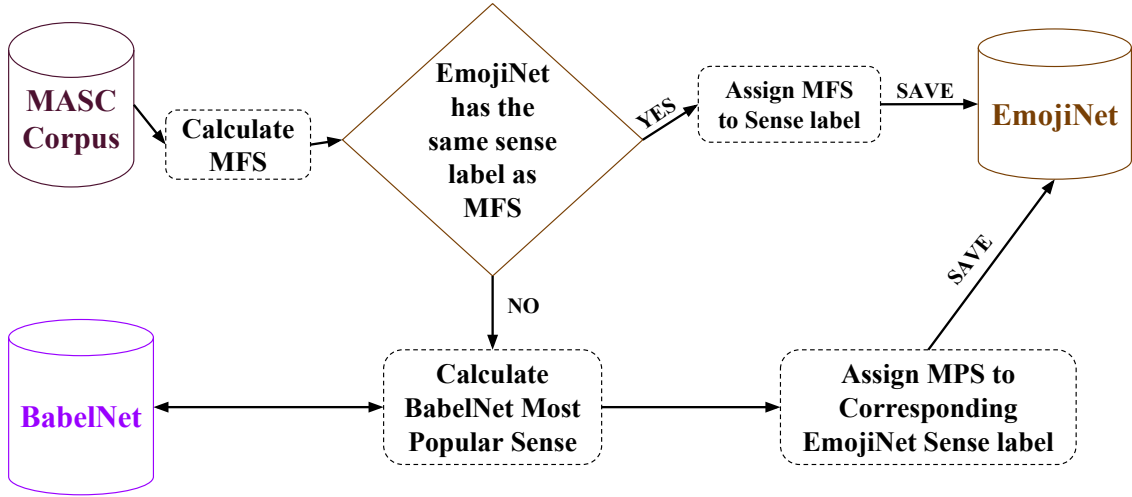


Figure 3.3: Assigning BabelNet Sense Definitions.

Figure 3.3 depicts the process of assigning BabelNet sense definitions to the sense labels in EmojiNet. We use the MASC annotations provided by Moro *et al.* to calculate the MFS of each word in the MASC corpus. Then, for all sense labels that are common for the MASC corpus and EmojiNet, we assign the calculated MFS for each of them as their corresponding sense definition and save their sense definitions in EmojiNet. However, not all sense labels in EmojiNet were assigned BabelNet senses in the above WSD task as several sense labels were not present in the MASC corpus. To assign sense definitions for those that were missed in the earlier WSD task, we defined a second WSD task based on the most popular sense (MPS) of each BabelNet sense. We define the MPS of a sense label as follows. For each BabelNet sense label B_s , we take the count of all sense definitions BabelNet lists for B_s . The MPS of B_s is the BabelNet sense ID that has the highest number of definitions for B_s . For sense labels that there are more than one MPS available in BabelNet, we manually assign the correct BabelNet sense ID. Once the MPS is calculated, those will be assigned to their corresponding sense labels in EmojiNet which were left out in the MFS-based WSD task.

3.5 Enhancing EmojiNet for Analysis Tasks

Making EmojiNet more useful beyond just serving as a machine-readable sense inventory and to enable its use for emoji analysis tasks on ill-formatted social media text required the following enhancements.

3.5.1 Adding Word Embeddings to EmojiNet

As pointed out earlier, the accuracy of the sense disambiguation tasks can be improved by incorporating more context words [Vasilescu et al. 2004], and NLP tools trained on well-formed text might not work well with the language variations seen in social media [Ritter et al. 2011]. Thus, to make EmojiNet a more robust tool for working with social media text processing, we derive additional context words based on word embedding models learned over Twitter and news articles, respectively.

We collected a Twitter dataset that contained emoji using the Twitter streaming API⁸ to train a word embedding model on tweets with emoji. The dataset was collected using emoji Unicodes as filtering words, over a four week period, starting from 6th August 2016 to 8th September 2016. It consists of 147 million tweets containing emoji. We removed all retweets and used the remaining 110 million unique tweets for training purposes. When training the Twitter-based word embedding model [Mikolov et al. 2013], we first convert all emoji into textual features using Emoji for Python⁹ API. Then we remove all stopwords and perform stemming across all tweets. We then feed all the training data (i.e. words found in tweets, including emoji) to the Word2Vec tool and train it using a Skip-gram model with negative sampling. We choose Skip-gram model with negative sampling to train our model as it is shown to generate robust word embedding models even when certain words are less frequent in the training corpus [Mikolov et al. 2013]. We set the number of dimensions of our model to 300 and the negative sampling rate to 10 sample words, which works well with medium-sized datasets [Mikolov et al. 2013]. We set the context word window to be 5 so that it will

⁸<https://dev.twitter.com/streaming/public>

⁹<https://pypi.python.org/pypi/emoji/>

consider 5 words to left and right of the target word at each iteration of the training process. This setting is suitable for sentences where average sentence length is less than 11 words, as is the case in tweets [Hu et al. 2013]. We ignore the words that occur fewer than 10 times in our Twitter dataset when training the word embedding model. We use a publicly available word embedding model that is trained over Google News corpus¹⁰ to learn additional context words for emoji sense definitions.

We follow an approach similar to the one presented by Eisner *et al.* when learning additional context words for emoji sense definitions [Eisner et al. 2016]. For each emoji $e_i \in E$, we extract the definition d_i of the emoji e_i and the set of all emoji sense definitions S_i of e_i from EmojiNet. Then, for each word w in d_i , we extract the twenty most similar words from the two word embedding models as two separate sets, namely $CW_{e_i}^T$ and $CW_{e_i}^N$. For example, for 🖱️, EmojiNet lists “A gun emoji, more precisely a pistol. A weapon that has potential to cause great harm” as its emoji definition. To generate context words, we replace each word in the definition above with the top twenty most similar words learned for it using the two word embedding models, respectively. We do the same for each emoji sense definition for 🖱️ as well. For each emoji sense definition $s_i \in S_i$ that belongs to e_i , we then extract the words w_{s_i} in $s_i \in S_i$ and repeat the same process to learn two separate context word sets $CW_{e_i-s_i}^T$ and $CW_{e_i-s_i}^N$, based on the twenty most similar words for each word w_{s_i} in $s_i \in S_i$. The separate sets allow us to mark if a context word was learned from social media (Twitter) or more well-formed text (news articles) in EmojiNet.

3.5.2 Adding Platform-specific Meanings to EmojiNet

As pointed out by [Miller et al. 2016], platform-specific emoji meanings could also play an important role in emoji understanding tasks. We came up with a list¹¹ of 40 most confused emoji (please refer to Appendix A for the full emoji list) based on the differences in their platform-specific images and crowd-provided senses, including the 25 emoji studied by Miller *et al.* We setup an Amazon

¹⁰<https://goo.gl/QaxjVC>

¹¹http://emojinet.knoesis.org/dataset/vendorspecific_emoji.html

Mechanical Turk (AMT) experiment to identify the platform-specific meanings associated with the 40 selected emoji. We selected five vendor platforms for our study, namely Google, Apple, Windows, Samsung, and Twitter, and extracted all of the emoji sense labels stored in EmojiNet. In our experiment, a single AMT task asks a worker to say whether they think that a given sense label is a reasonable sense for a platform-specific emoji. Radio buttons (agree or disagree) are used to record their decisions, along with a text field to give a brief explanation. Results with no, repeating, or nonsense explanations were filtered away under the assumption that the worker was a spammer. We conducted a total of 14,448 such AMT tasks, of which 1,128 were filtered as spam.

We looked at what were the emoji that had platform specific meanings. We specifically look for meanings that are unique for certain emoji when they were shown in certain platforms. We found 27 emoji (67.50%) that had platform specific meanings. For example, for 🍷, we noticed that windows platform was the only one that shows a smiling face with teeth displaying as shown in 😁. Therefore, the AMT workers have assigned laugh as a meaning for 😁 but not for any other emoji depictions for the same Unicode representation including 🍷. We also noticed that the Samsung platform-based emoji had the least number of meanings associated with their images, which tells us Amazon workers had hard time agreeing with each other on the emoji meanings when Samsung images are displayed. Google platform images had the highest agreement among emoji senses. We've added these vendor-specific meanings into EmojiNet dataset.

3.6 EmojiNet Web application and REST API

To make it easy to browse and programmatically access the EmojiNet dataset, we host EmojiNet as a web application at <http://emojinet.knoesis.org/>. The web application supports searching the EmojiNet dataset based on the Unicode character representation, name, or the short code of an emoji. It also supports browsing EmojiNet by specifying the part-of-speech tags of the emoji senses (see Figure 3.4). A REST API is implemented so that computer applications can also access the

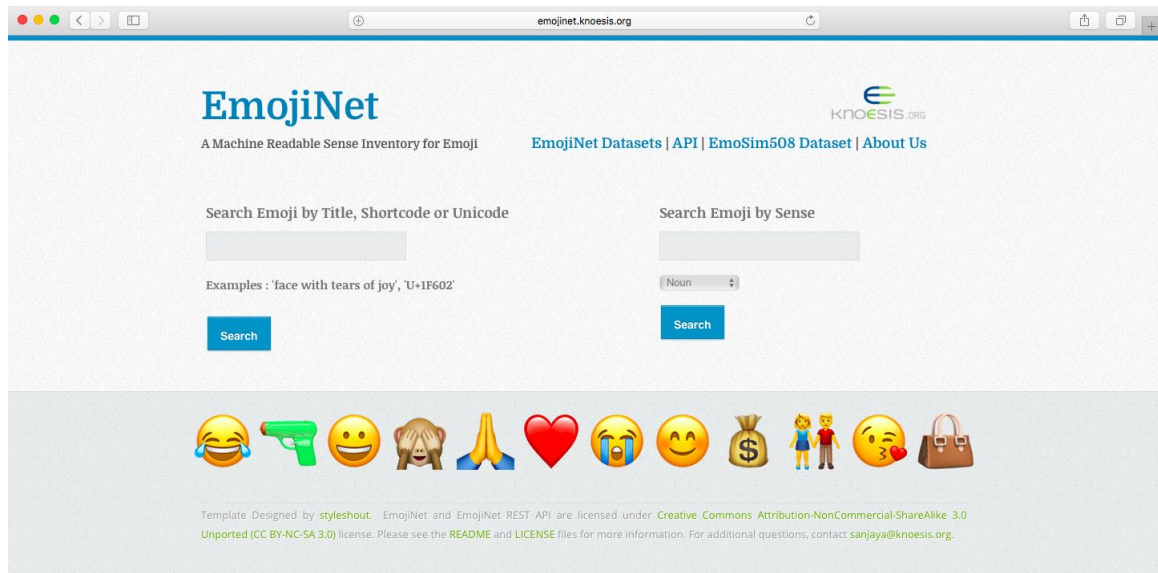


Figure 3.4: EmojiNet Web Application at <http://emojinet.knoesis.org/>.

EmojiNet dataset. The API has a series of methods that can be invoked over an HTTP connection that return data in a JSON object format. The resource, along with the complementary sense embedding, vendor-specific sense data and REST API with documentation, is freely available to the public for research use. Each method in the EmojiNet REST API is discussed in Appendix B by providing the functionality of the method, method signature, URL parameters of the method, and a sample request that shows how the method is invoked. For a detailed description of each method signature including sample responses and error codes, please refer to the online EmojiNet REST API available at - <http://emojinet.knoesis.org/api.php>.

Table 3.2 lists some summary statistics for the data stored in EmojiNet and the emoji data distribution. Each emoji in EmojiNet carries all features listed in Table 3.2 except related emoji. A total of 7,544 related emoji pairs have been stored in EmojiNet that belongs to 1,755 emoji. There are 6,057 emoji keywords, 18,615 images, and 141 categories shared across the emoji. An emoji in EmojiNet has 5 to 6 senses on average.

Figure 3.5 plots the number of unique emoji senses for each emoji stored in EmojiNet. A selected set of emoji are also shown there. Emoji with a diverse set of senses include 🍌 (55) followed by 🍋

Table 3.2: EmojiNet Statistics.

Emoji feature	# of emoji with each feature	# of data stored for each feature
Unicodes	2,389	2,389
Emoji Names	2,389	2,389
Short Codes	2,389	2,389
Descriptions	2,389	2,389
Keywords	2,389	6,057
Images	2,389	18,615
Related Emoji	1,755	7,544
Categories	2,389	141
Senses	2,389	12,904

(49). For example, 🍌 had senses ranging from chocolate to smelly. When looking at the senses, it was evident that most of the senses are based on the look and feel of the emoji. For example, 💩 had many sense variations that interpret as feces, and some sense variations which were based on the color and the shape of the emoji. 💧 had senses ranging from sweat to rain. We also noticed 😍, 😘 and 😂, which are three of the most popular emoji on Twitter¹², were among EmojiNet's top 10 emoji with most number of senses. We examined the emoji that had least number of emoji senses (1). Those include blood type emoji such as 🇦 and 🇦🇧, buttons such as 🇨🇴🇴🇱 and 🇫🇷🇪🇪, and newly introduced emoji such as 🦌 and 🧑. We found that all of them do not exist in The Emoji Dictionary website, hence they did not have any crowd-generated emoji meanings saved in EmojiNet. We also noticed that they have only one keyword listed in the Unicode Consortium website as the intended meaning. Some of them, such as animal faces, were recently introduced.

¹²<http://www.emojitracker.com/>

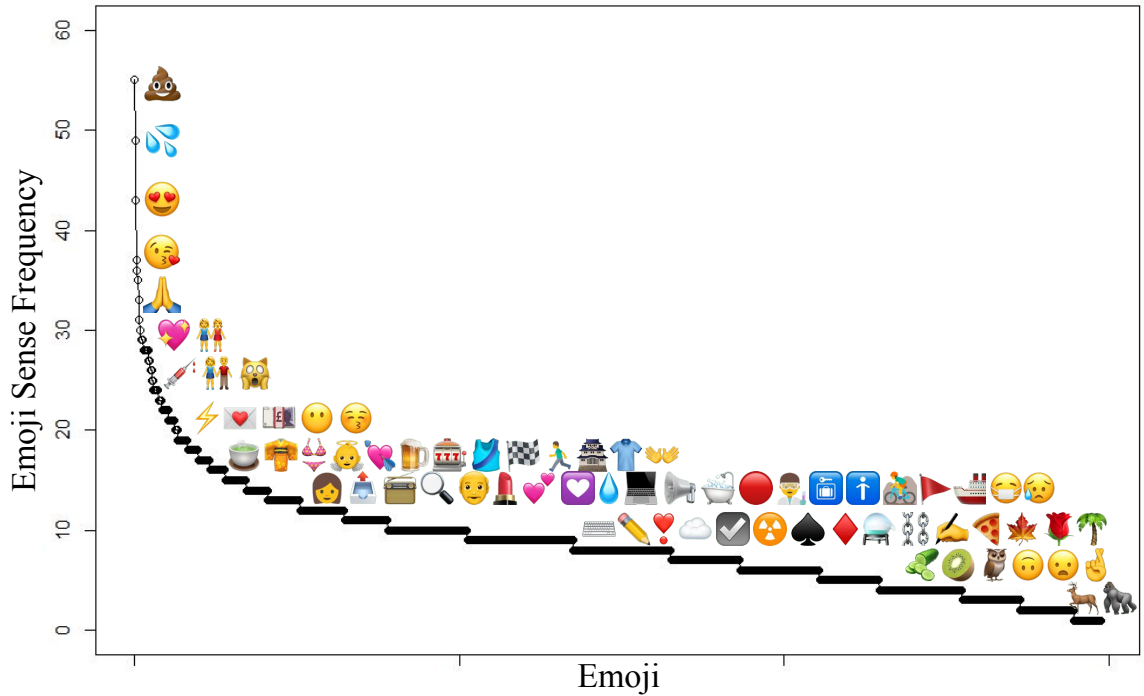


Figure 3.5: Emoji Sense Distribution.





3.7 Dataset Evaluation

This section evaluates the process we used to curate the data published in EmojiNet. In particular, we evaluate the nearest neighborhood-based image processing algorithm that we used to integrate emoji resources and the most frequent sense-based and most popular sense-based WSD algorithms that we used to assign meanings to emoji sense labels.

3.7.1 Resource Integration Evaluation

We evaluate how the nearest neighborhood-based image processing algorithm matches images from The Emoji Dictionary website (i.e., I_t) with the images downloaded from the Unicode Consortium website (i.e., I_x). The Unicode Consortium website contains multiple vendor-specific images for a given emoji that depict how an emoji looks on those vendors' platforms (i.e., different emoji for different platforms such as Apple, Samsung, Windows, Twitter etc.). Since we have indexed all

vendor-specific images of each emoji under that emoji’s Unicode representation, to correctly map an image in I_t with I_x , we only require an image in I_t to match with any one of those vendor-specific images in I_x for a given emoji. Once the matching process is completed, we pick the top ranked match for each emoji based on the dissimilarity of the two matched images and manually evaluate them for equality. The Unicode representation of the top ranked matched image from I_x will be assigned as the Unicode representation of the matching image from I_t . Though the image processing algorithm we used is naive, it works well for our study as the images of the emoji are not complex (i.e., each image has one object such as a face or a fruit) and they do not contain complex color combinations (i.e., one or two colors with a transparent/white background). The image processing algorithm we used combines color (spectral) information with spatial (position/distribution) information and tends to represent those features well when the images are simple.

Out of the 1,074 image instances we checked, our algorithm could correctly find matching images for 1,034 images in I_t with an accuracy of 96.27%. An error analysis performed on the 40 incorrect matches revealed that 13 family emoji, 9 person/family emoji, and 8 clock emoji were identified incorrectly among others. These image types had minimal differences in either objects or color, hence the algorithm had failed to match them correctly. For example, 9 person/family emoji had very slight differences in objects such as long hair in one image versus the short hair in the other (e.g.,  Vs ). The clock emoji had their arms at different locations while the color of the images were identical (e.g.,  Vs ). We manually corrected the 40 incorrect matches and assigned them their correct Unicode representations to complete the integration between The Emoji Dictionary data with Unicode Consortium data.

3.7.2 Sense Assignment Evaluation

Here we discuss how we evaluated the most frequent sense-based and most popular sense-based word sense disambiguation algorithms that we used to link emoji sense labels with BabelNet sense definitions. To do this evaluation, we sought the help of two human judges who have experience in

Table 3.3: Word Sense Disambiguation Statistics.

	Correct	Incorrect	Total
Noun	6,633 (86.64%)	1,022 (13.36%)	7,655
Verb	2,231 (77.14%)	661 (22.86%)	2,892
Adjective	1,915 (81.24%)	442 (18.76%)	2,357
Total	10,779 (83.53%)	2,125 (16.47%)	12,904

NLP research. We provided them with all emoji included in EmojiNet, listing all the valid sense labels for each emoji and their corresponding BabelNet senses (BabelNet sense IDs with definitions) calculated though either MFS or MPS baselines. They were asked to mark whether they thought that the suggested BabelNet sense ID was the correct sense ID for the emoji sense label listed. We calculated the agreement between the two judges for this task using Cohen’s kappa coefficient¹³ and obtained an agreement value of 0.7134, which is considered to be a good agreement.

Out of the 12,904 sense labels we provided them to disambiguate, 7,815 appeared in both the EmojiNet dataset and MASC dataset, so they were assigned BabelNet sense definitions through the MFS-based WSD approach. Our judges evaluated the sense assignments based on whether they thought that the suggested BabelNet sense ID assigned by the MFS baseline was the correct sense ID for the emoji sense label. They decided that 6,673 sense labels were assigned correct BabelNet sense IDs, yielding an accuracy of 85.38% for the MFS baseline. Judges then assigned the correct sense IDs for the 1,142 sense labels that were sense disambiguated incorrectly by the MFS baseline. The remaining 5,089 sense labels which were not assigned senses by the MFS baseline were considered for a second WSD task based on a MPS baseline. While evaluating the accuracy of the MPS baseline, the judges followed the same approach that they followed for evaluating the MFS baseline. Based on the evaluation results, we found that the MPS baseline has achieved 80.68% accuracy in the WSD task. There were 983 sense labels which were sense disambiguated incorrectly

¹³<https://goo.gl/szv50P>

3.8. EMOJINET AT WORK - AN EXPERIMENT ON CALCULATING EMOJI SIMILARITY⁴³

in this approach, which were then corrected by the judges. Overall, the two WSD baselines have correctly sense disambiguated a total of 10,779 sense labels, yielding an accuracy of 83.53% for the WSD task. Table 3.3 integrates the results obtained by both word sense disambiguation algorithms for different part-of-speech tags. The results shows that the two WSD approaches we used have performed reasonably well in disambiguating the sense labels in EmojiNet.

3.8 EmojiNet at Work - An Experiment on Calculating Emoji Similarity

Similar to semantic similarity of words¹⁴, we define emoji similarity based on the likeness of their meaning as defined by the sense labels assigned to each emoji. This is a new notion of ‘emoji similarity’ compared to previous work that defined similarity by emoji functionality or topic [Barbieri et al. 2016; Eisner et al. 2016] and is uniquely enabled by EmojiNet’s sense repository. This sense similarity is similar to how semantic similarity measures have been defined using sense inventories such as WordNet¹⁵ and BabelNet. Since EmojiNet carries functional and topical emoji meanings available in the Unicode Consortium and The Emoji Dictionary websites in addition to the other intended meanings of emoji if any, our method complements other similarity measures. Next, we describe a use-case where we model an emoji similarity graph using the emoji present in EmoTwi50 [Barbieri et al. 2016] dataset created by Barbieri *et al.* to explain sense-based emoji similarity.

EmoTwi50 is a dataset that contains 25 manually created and 25 randomly created emoji pairs, totaling 100 unique emoji. Barbieri *et al.* created and used this dataset to find functional and topical similarity of the 50 emoji pairs. We use it to create our emoji similarity graph based on emoji senses. We first extract the sense labels of the 100 emoji in EmoTwi50 dataset from EmojiNet. A node in the emoji similarity graph is an emoji and an edge exists if there is at least one common

¹⁴<https://goo.gl/ITXkAT>

¹⁵<https://wordnet.princeton.edu/>

3.8. EMOJINET AT WORK - AN EXPERIMENT ON CALCULATING EMOJI SIMILARITY44

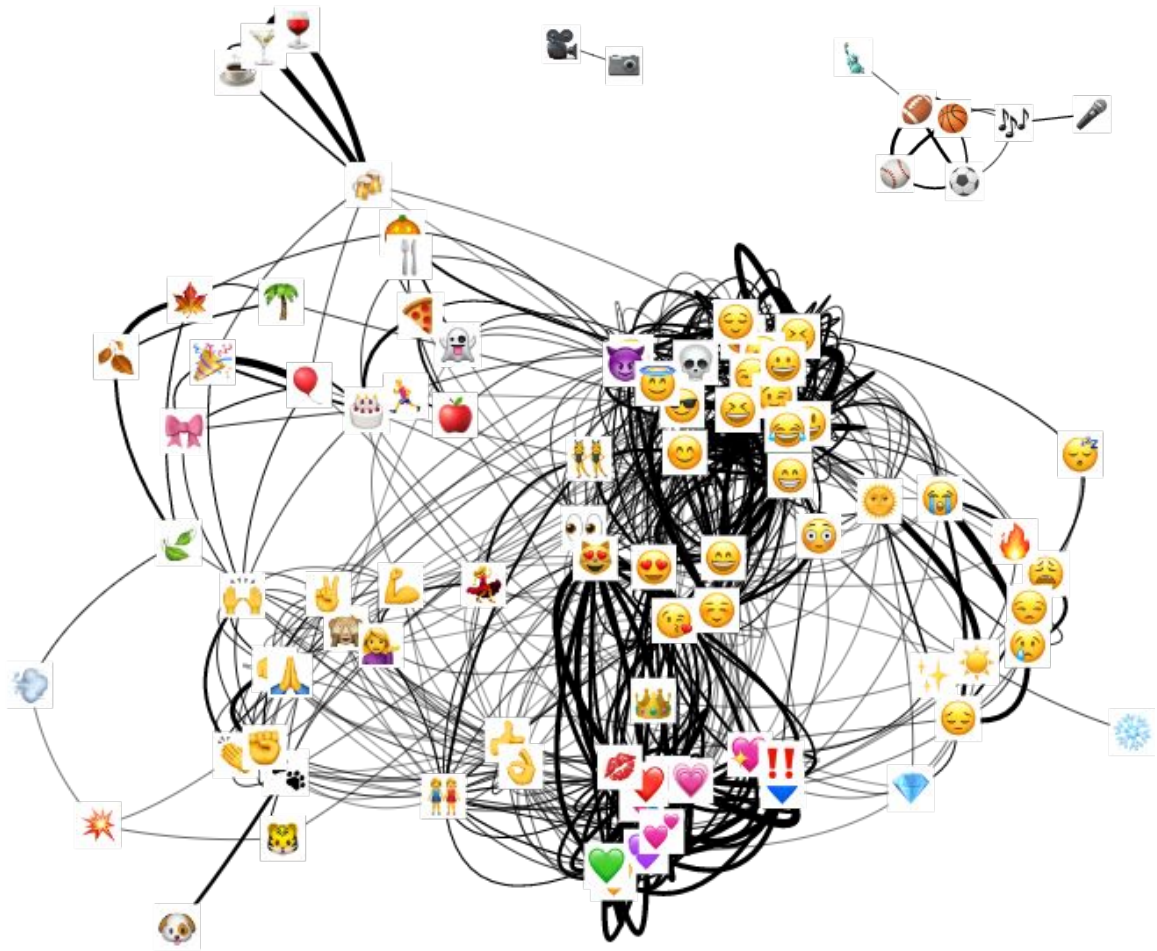


Figure 3.6: Emoji Clusters using Emoji Sense Overlap.

sense between them. Figure 3.6 visualizes this graph, with the thickness of an edge corresponding to the number of shared emoji senses between them. We then run a label propagation community detection algorithm [Barber and Clark 2009] to identify emoji communities (clusters) based on their sense overlap. This revealed 16 clusters in our graph, each of which represents ‘sense-similar’ emoji. We have list a selected set of emoji within different clusters and label them in Table 3.4. We can see that the smiling face emoji have been clustered together while sad faces, hearts, drinks, and hand symbols form their own clusters. We also notice two islands, which we have labeled as cameras and sports & entertainment.

Table 3.4: Selected Emoji Sense Clusters in EmoTwi50.

Cluster Name	Emoji
Smiling Faces	
Love	
Sad Faces	
Drinks	
Cameras	
Sports & Entertainment	




















Once the graph is computed, we can use any traditional semantic, path or set similarity measure to find the sense similarity between any two emoji in the graph. For example, we use Jaccard Similarity¹⁶ to find the sense similarity between 😊 and 😄. Both emoji have 12 sense labels each, shares 9 sense labels, and have 15 unique sense labels between them. Thus, the sense similarity between them can be calculated as the ratio between 9 and 15, which gives 0.60. Table 3.5 lists the ten most similar emoji pairs calculated based on EmoTwi50 dataset. We can replace Jaccard Similarity with a sophisticated similarity measure to improve the results shown. The emoji similarity dataset we created using Jaccard Similarity is available to download at <http://emojinet.knoesis.org/>.

3.9 Summary

This chapter presented the details of the creation of EmojiNet, the largest machine-readable emoji sense inventory that links Unicode emoji representations to their English meanings extracted from the Web. We described how (i) three open resources were integrated to build EmojiNet, (ii) word embedding models and vendor-specific emoji senses were used to improve EmojiNet, and (iii) how the resource building process was evaluated. We also discussed the web application we developed

¹⁶<https://goo.gl/RKkRzF>

Table 3.5: Ten Most Similar Emoji Pairs Based on Jaccard Similarity.

Emoji Pair	Similarity
 	0.60
 	0.57
 	0.56
 	0.52
 	0.52
 	0.50
 	0.50
 	0.50
 	0.48
 	0.47

to browse EmojiNet dataset, the REST API to access the functionality of EmojiNet and showed how EmojiNet can be used to calculate emoji sense similarity. We will discuss how EmojiNet can be used to solve emoji understanding tasks in the next two chapters of this dissertation.

4

Applications of EmojiNet – Emoji Similarity

4.1 Overview


In this chapter, we look at how EmojiNet can be used to solve emoji similarity tasks. First, we present a comprehensive study on measuring the semantic similarity of emoji using emoji embedding models. We extract machine-readable emoji meanings from EmojiNet to model the meaning of an emoji. Using pre-trained word embedding models learned over a Twitter dataset of 110 million tweets and a Google News text corpus of 100 billion words, we encode the extracted emoji meanings to obtain emoji embedding models. To create a gold standard dataset for evaluating how well the emoji embeddings measure similarity, we ask ten human annotators who are knowledgeable about emoji to manually rate the similarity of 508 pairs of emoji. This dataset of human annotations, which we call ‘EmoSim508’, is made available with this dissertation for use by other researchers. We evaluate the emoji embeddings by first establishing that the similarity measured by our embedding models align with the ratings of the human annotators using statistical measures. Then, we apply our emoji embedding models to a sentiment analysis task to demonstrate the utility of them in a

real-world NLP application. Our models were able to correctly predict the sentiment class of tweets laden with emoji from a benchmark dataset [Novak et al. 2015] with an accuracy of 63.6 (7.73% improvement), outperforming the previous best results on the same dataset [Barbieri et al. 2016; Eisner et al. 2016]¹.

4.2 Representation of Emoji Meaning

EmojiNet provides different types of information on emoji which one can use to train computer models. Here, we are interested in the information in EmojiNet that we can use to represent the meaning of an emoji. We consider three different ways to represent the meaning of an emoji using the information in EmojiNet. Specifically, we extract emoji descriptions, emoji sense labels, and the emoji sense definitions of each emoji sense from EmojiNet to model the meaning of an emoji. We discuss each briefly below.


4.2.1 Emoji Description (*Sense_Desc.*)

Emoji descriptions give an overview of what is depicted in an emoji and its intended uses. For example, for the pistol emoji , EmojiNet lists “*A gun emoji, more precisely a pistol. A weapon that has potential to cause great harm. Displayed facing right-to-left on all platforms*” as its description. One could use this information to get an understanding of how the pistol emoji should be used in a message.


4.2.2 Emoji Sense Labels (*Sense_Label*)

Emoji sense labels are word-POS tag pairs (such as `laugh(noun)`) that describe the senses and their part-of-speech under which an emoji can be used in a sentence. Emoji sense labels can act as words that convey the meaning of an emoji and thus, can play an important role in understanding

¹Content presented in this chapter were previously published in [Wijeratne et al. 2017b]

the meaning of an emoji. For example, for pistol emoji , EmojiNet lists 12 emoji sense labels consisting of 6 nouns (gun, weapon, pistol, violence, revolver, handgun), 3 verbs (shoot, gun, pistol) and 3 adjectives (deadly, violent, deathly).

4.2.3 Emoji Sense Definitions (*Sense_Def.*)

Emoji sense definitions are the textual descriptions that explain each sense label and how those sense labels should be used in a sentence. For example, for the gun(Noun) sense label of the pistol emoji , EmojiNet lists 5 sense definitions that complement each other². These emoji sense definitions can be valuable in understanding the meaning of an emoji; thus, we use them to represent the meaning of an emoji.

4.2.4 Learning the Emoji Embedding Models

Once the machine-readable emoji descriptions are extracted from EmojiNet, we use word embedding models [Mikolov et al. 2013] to convert them into a vectorial representation. A word embedding model is a neural network that learns rich representations of words in a text corpus. It takes data from a large, n -dimensional ‘word space’ (where n is the number of unique words in a corpus) and learns a transformation of the data into a lower k -dimensional space of real numbers. This transformation is developed in a way that similarities between the k -dimensional vector representation of two words reflects semantic relationships among the words themselves. Word embedding models are inspired by the distributional hypothesis (i.e., words that are co-occurring in the same contexts tend to carry similar meanings), hence the semantic relationships among word vectors are learned based on the word co-occurrence in contexts (e.g., sentences) extracted from large text corpora. Mikolov *et al.* have shown that these word embeddings can learn different types of semantic relationships, including gender relationships (e.g., King-Queen) and class inclusion (e.g., Clothing-Shirt) among many others [Mikolov et al. 2013]. Similar to word embedding models,

²<https://goo.gl/gm7TQ2>

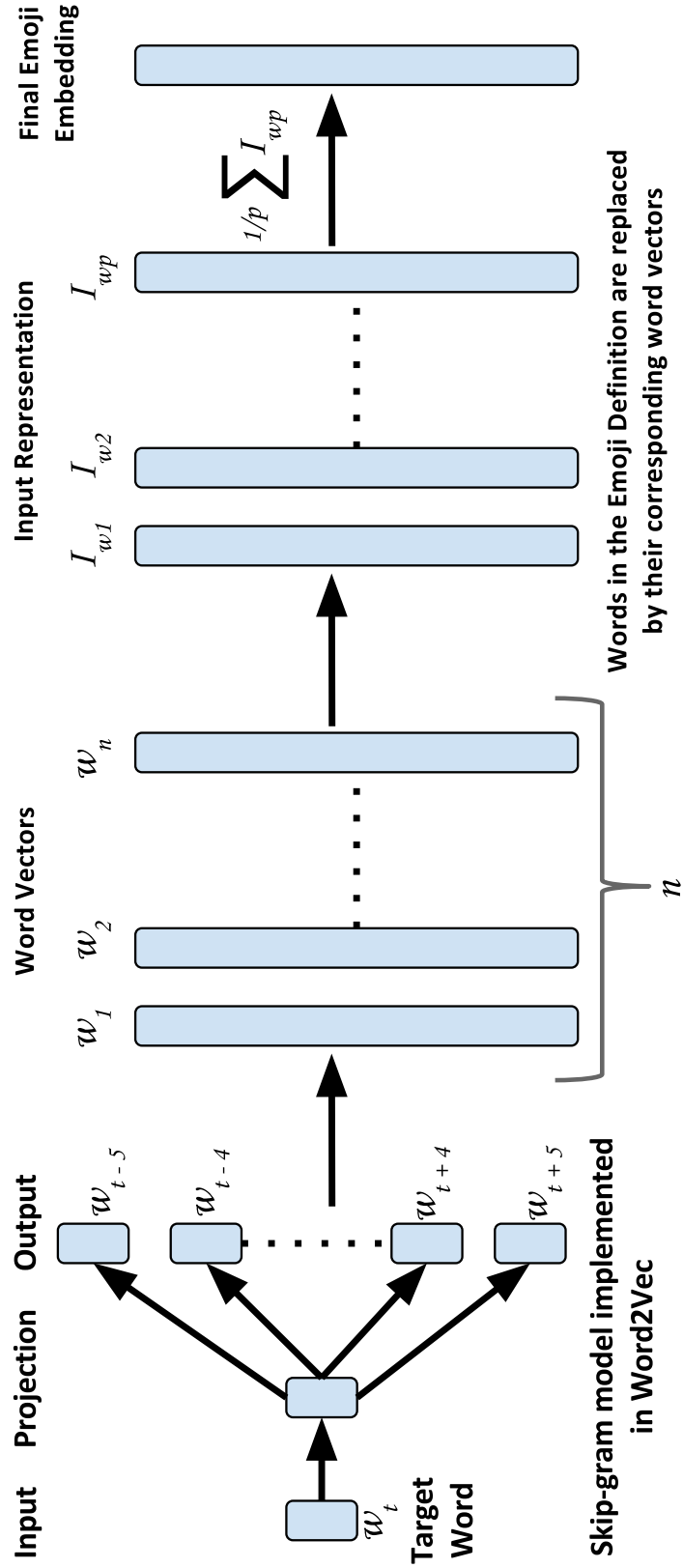


Figure 4.1: Learning Emoji Embedding Models using Word Vectors.

an emoji embedding model is defined as an emoji symbol and its learned k -dimensional vector representation.

We chose two different types of datasets, namely, a tweet corpus and a Google News corpus, to train emoji embedding models. We made this selection to make it easy to compare our emoji embedding models with other works that have used embedding models based on tweet text and Google News text. To train the Twitter word embedding model, we first collected a Twitter dataset that contained emoji using the Twitter public streaming API³. The dataset was collected using emoji Unicodes as filters over a four week period, from August 6th, 2016 to September 8th, 2016. It consists of 147 million tweets containing emoji. We first removed all retweets and then converted all emoji in the remaining 110 million unique tweets into textual features using the Emoji for Python⁴ API. The tweets were then stemmed before being processed with Word2Vec [Mikolov et al. 2013] using a Skip-gram model with negative sampling. This process is depicted in Figure 4.1. We choose the Skip-gram model with negative sampling to train our model as it is shown to generate robust word embedding models even when certain words are less frequent in the training corpus [Mikolov et al. 2013]. We set the number of dimensions of our model to 300 and the negative sampling rate to 10 sample words, which are shown to work well empirically [Mikolov et al. 2013]. We set the context word window to 5 (words w_{t-5} to w_{t+5} in Figure 4.1) so that it will consider 5 words to left and right of the target word (word w_t in Figure 4.1) at each iteration of the training process. This setting is suitable for sentences where the average sentence length is less than 11 words, as is the case in tweets [Hu et al. 2013]. We ignore the words that occur fewer than 10 times in our Twitter dataset when training the word embedding model. We use a publicly available word embedding model that is trained over Google News corpus⁵ to obtain Google News word embeddings.

We use the learned word vectors to represent the different types of emoji definitions listed in Section 4.2. All words in each emoji definition are replaced with their corresponding word vectors

³<https://dev.twitter.com/streaming/public>

⁴<https://pypi.python.org/pypi/emoji/>

⁵<https://goo.gl/QaxjVC>

as shown in Figure 4.1. For example, all words in the pistol emoji’s 🖱️ description, which is “*A gun emoji, more precisely a pistol. A weapon that has potential to cause great harm. Displayed facing right-to-left on all platforms*” are replaced by the word vectors learned for each word. Then, to get the emoji embedding, the word vectors of all words in the emoji definition are averaged into form a final single vector of size 300 (the dimension size). The vector mean (or average) adjusts for word embedding bias that could take place due to certain emoji definitions having considerably more words than others [Kenter et al. 2016]. If the total number of words in the emoji definition is p , the combined word vector \mathbf{V}_p is calculated by:

$$\mathbf{V}_p = 1/p \sum_{i=0}^p \mathbf{w}_i$$

Using the three different emoji definitions and two types of word vectors learned over Twitter and the Google News corpora, we learn six different embeddings for each emoji. Then we integrate all words in the three types of emoji definitions into a set called (*Sense_All*) and learn two more emoji embeddings over them by using the two types of word vectors. We take this step as prior research suggests that having more words in an emoji definition could improve the embeddings learned over them [Eisner et al. 2016; Pohl et al. 2017]. Thus, we learn a total of 8 embeddings for emoji. The utility of each embedding as a similarity measure is discussed next.

4.3 Ground Truth Data Curation

Once the emoji embedding vectors are learned, it is necessary to evaluate how well those represent emoji meanings. For this purpose, we create an emoji similarity dataset called ‘EmoSim508’ that consist of 508 emoji pairs which were assigned similarity scores by ten human judges. This section discusses the development of the EmoSim508 emoji similarity dataset, which is available at <http://emojinet.knoesis.org/emosim508.php>.

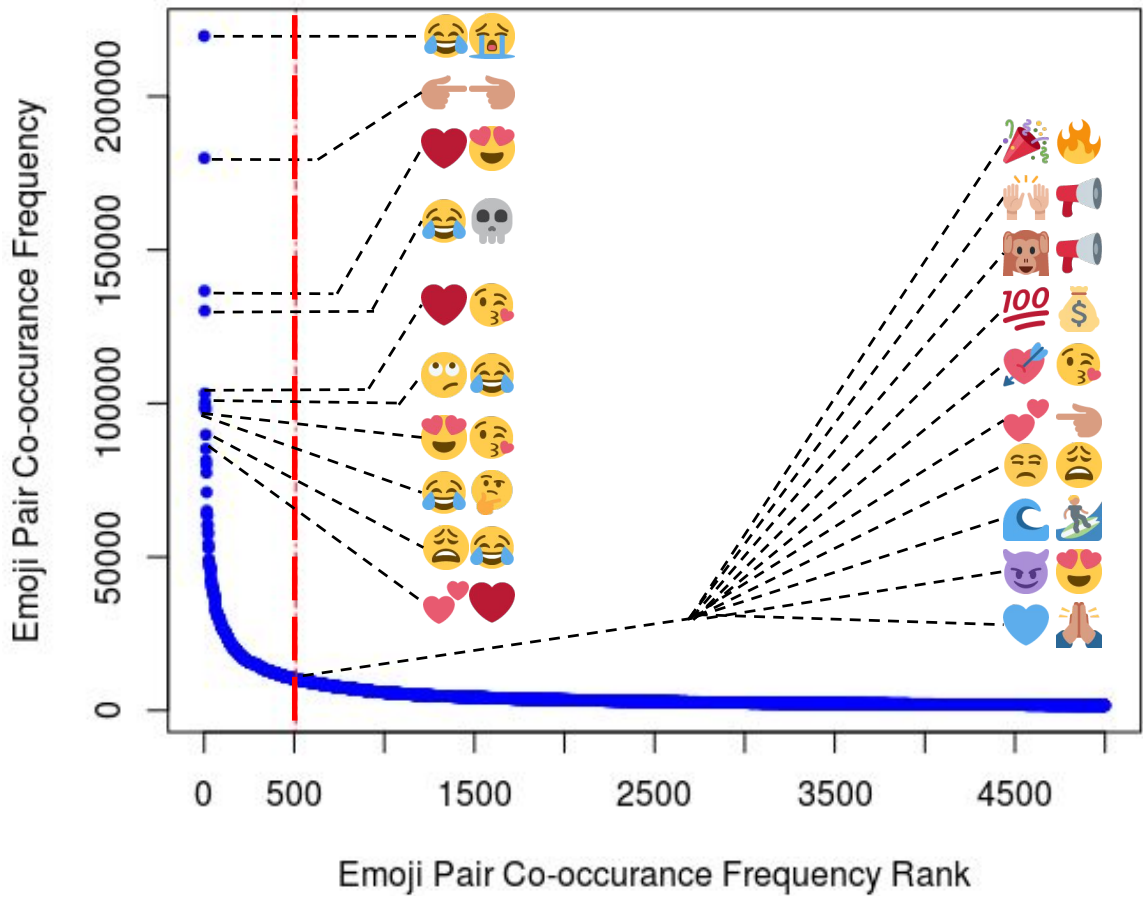


Figure 4.2: Emoji Co-Occurrence Frequency Graph.

4.3.1 Emoji Pair Selection

Curating a reasonable sample of emoji pairs for human evaluation is a critical step: there are 2,389 emoji, leading to over 5 million emoji pairs, which would be impossible to ask a human to evaluate for their similarity. Hand-picking emoji pairs might not be a good approach as such a dataset would not cover a wide variety of similarities or could be biased towards certain relationships that commonly exist among emoji [Barbieri et al. 2016]. Furthermore, random sampling of the emoji pairs will lead to many unrelated emoji as suggested by Barbieri *et al.* [Barbieri et al. 2016], making the dataset less useful as a gold standard dataset. We thus sought to curate the EmoSim508 dataset in such a way that the emoji pairs in it are not hand-picked but still represent a ‘meaningful’ dataset. By meaningful, we mean that the dataset contains emoji pairs that are often seen together in practice. The dataset should also have pairs that are related, unrelated, and the shades in-between, leading to a diverse collection of examples for evaluating a similarity measure. To address this, we consider the most frequently co-occurring emoji pairs from the Twitter corpus used to learn word vectors in Section 4.2.4 and created a plot of how often pairs of emoji co-occur with each other. From this plot, shown in Figure 4.2, we select the top-k emoji that cover 25% of our Twitter dataset (shown in the dotted red line in Figure 4.2). This resulted in the top 508 emoji pairs. Since the co-occurrence frequency plot is decidedly heavy-tailed (the blue line), we chose the 25% threshold, giving us a dataset which is 10 times bigger than the previous dataset used by Barbieri *et al.* [Barbieri et al. 2016] to calculate emoji similarity. These 508 emoji pairs have 158 unique emoji. We have also shown the top 10 and bottom 10 emoji pairs based on their co-occurrence frequency in Figure 4.2. We can observe that the face emoji are dominant in the top 10 emoji pairs while bottom 10 contain few interesting emoji pairs such as 📧 and 🙄, 🙄 and 🙄, and 🙄, and 🌊 and 🏊.

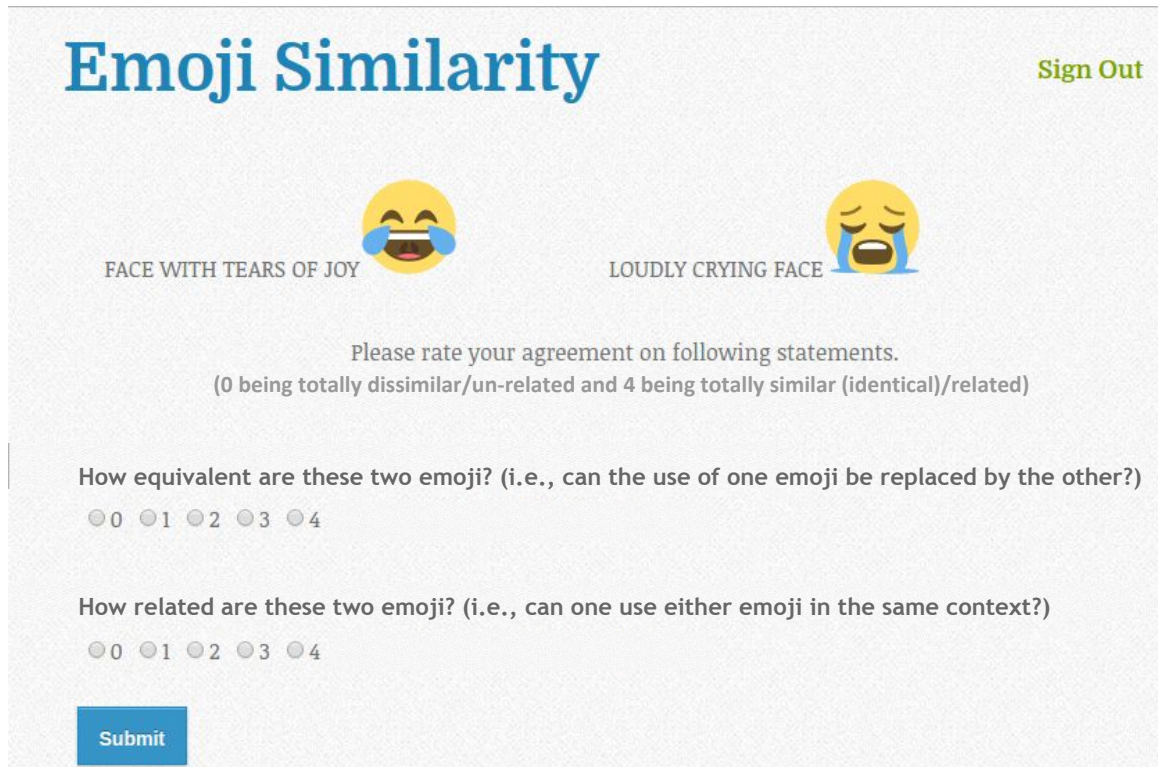
4.3.2 Human Annotation Task

We use human annotators to assign similarity scores to each emoji pair in the EmoSim508 dataset. A total of ten annotators were used, all of whom were graduate students at Wright State University, and of whom four were male and six were female. Their ages ranged from 24 years to 32 years; past studies suggest people within this age range use emoji frequently⁶. The annotators were shown a webpage (a screen shot is shown in Figure 4.3) with two emoji and were prompted with two questions, one related to emoji equivalence and the other related to emoji relatedness, which they were required to answer on a five-point Likert scale [Likert 1932] ranging from 0 to 4, where 0 means the emoji were dissimilar and 4 means the emoji were identical. We selected the five-point Likert scale for our study for two main reasons. Firstly, past research has shown that Likert scale is best suited for questionnaire-based user studies and five-point scale have shown to perform better than other scales (seven-points and ten-points) empirically [Revilla et al. 2014]. Secondly, many human annotators-involved word similarity experiments have used the same Likert scale from 0 to 4 to calculate the similarity of words [Snow et al. 2008]. The two questions we asked from the annotators were:

- **Q1.** How equivalent are these two emoji?
(i.e., can the use of one emoji be replaced by the other?)
- **Q2.** How related are these two emoji?
(i.e., can one use either emoji in the same context?)

We asked Q1 to understand whether an equivalence relationship exists between an emoji pair and Q2, to understand whether a relatedness relationship exists between them. Annotators answered the same two questions for all 508 emoji pairs in the EmoSim508 dataset. We then averaged values received as answers for the ordinal selections (0 to 4) for both questions separately and assign the emoji pair an emoji equivalence score and an emoji relatedness score. Then we average the two

⁶<https://goo.gl/GSbcGL>



Emoji Similarity Sign Out

FACE WITH TEARS OF JOY 😂 LOUDLY CRYING FACE 😭

Please rate your agreement on following statements.
(0 being totally dissimilar/un-related and 4 being totally similar (identical)/related)

How equivalent are these two emoji? (i.e., can the use of one emoji be replaced by the other?)
 0 1 2 3 4

How related are these two emoji? (i.e., can one use either emoji in the same context?)
 0 1 2 3 4

Figure 4.3: A Screen shot of the Web Application Used for the Annotation Task.

values to calculate the final emoji similarity score for a given pair of emoji. We use the final emoji similarity score to evaluate our emoji embedding models.

4.3.3 Annotation Evaluation

We conducted a series of statistical tests to verify that EmoSim508 is a reliable dataset, that is, to ensure that the annotators did not randomly answer the task’s questions [Artstein and Poesio 2008]. To verify this, we measured the inter-annotator agreement. Since we had ten annotators who used ordinal data to evaluate the similarity of emoji, we selected Krippendorff’s alpha coefficient α to calculate the agreement among annotators [Hayes and Krippendorff 2007]. We calculated annotator agreement for each question separately and observed an α value of 0.632 for Q1 and an α value of 0.567 for Q2. This tells us that the emoji similarity evaluation was not an easy task for the annotators and their agreement is slightly better when deciding on two emoji pairs for equivalence

than relatedness. In our dataset, a lot of annotators have agreed on the non-equivalence of emoji pairs, thus, we believe that the slightly higher α score for agreeing on the equivalence of an emoji pair could be a result of that.

To evaluate how reasonable are the scores provided by the human annotators, we look at the emoji pairs with highest inter-annotator agreement for each ordinal value in the Lickert scale (0 to 4) in Figure 4.4. Here, we focus on annotator agreement at each level of the Lickert scale (0 to 4). We notice that all annotators have agreed that the 🎵 and 🎶 emoji show an equivalence relationship. All other emoji pairs shown for ordinal value 4, which stands for ‘equivalent or fully related’, show high agreement (a minimum of 8/10) among the annotators. Ordinal value 3, which stands for ‘highly similar or closely related’, show medium agreement (a minimum of 5/10) among annotators. Ordinal values 1 and 2, standing for ‘slightly similar or slightly related’ and ‘similar or related’, respectively, also show medium agreement (a minimum of 5/10) among the top-5 emoji pairs for each ordinal value. Finally, ordinal value 0, which stands for ‘dissimilar or unrelated’, show full agreement (10/10) among annotators for a total of 184 emoji pairs. The annotators have unanimously agreed that there is no relatedness and equivalence relationships exist for a group 31 and 153 emoji pairs, respectively. This further shows that it has been easier for them to agree on the dissimilarity of a pair of emoji than on its similarity or relatedness.

Figure 4.5 depicts the distribution of the mean of the annotator ratings (line plot) and one standard deviation from the mean (ribbon plot) for each emoji pair for each question. For both questions, the mean of each plot shows a near-linear trend, proving that our dataset captures diverse types of relationships. Specifically, for question 1, we find a near-linear trend in the mean distribution for emoji pairs where the mean user rating is between 0.66 and 4. For question 2, we find a similar trend for emoji pairs where the mean rating is between 1 and 4. For both questions, the deviation bands are dense, especially in the range of 0.75 – 2.5, which is to be expected. We also note that the deviation does not span beyond one rating (e.g., the deviation bands at a mean of 2 tend to span between 1 and 3). This reasonable deviation further speaks for the diversity of responses. The size








































































































Ordinal Rating	0		1		2		3		4		
	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	
Emoji Pairs with Highest Agreement	 		 	 	 	 	 	 	 	 	 
			 	 	 	 	 	 	 	 	 
			 	 	 	 	 	 	 	 	 
	 	 	 	 	 	 	 	 	 	 	 
			 	 	 	 	 	 	 	 	 

Figure 4.4: Top-5 Emoji Pairs with Highest Inter-annotator Agreement for Each Ordinal Value from 0 to 4.

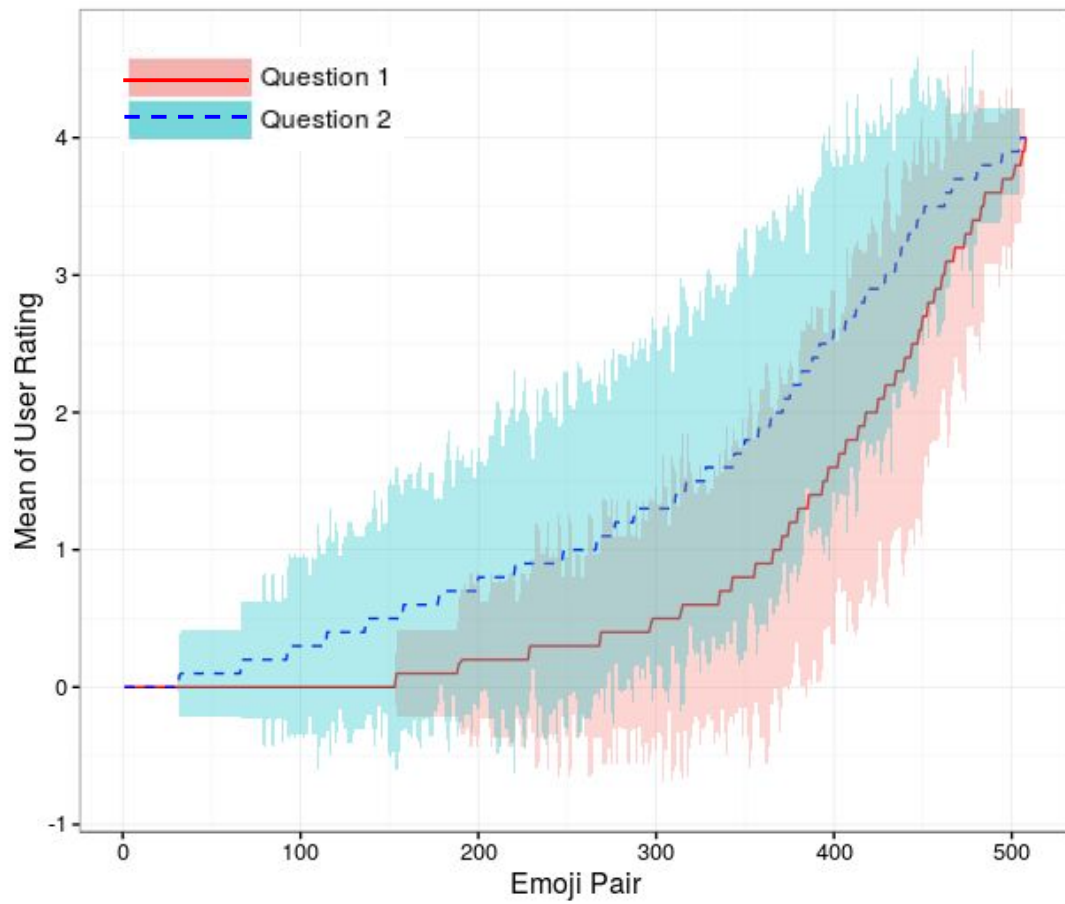


Figure 4.5: Distribution of the Mean of User Ratings.

of these deviation bands decrease as we approach extreme values (i.e., emoji definitely similar and definitely different). We notice an elbow (from $(0, 0)$ to $(153, 0)$) at the start of the mean distribution for Q1 due to the strong agreement among annotators for the unrelated emoji. This shows that even though we selected highly co-occurring emoji pairs from a Twitter corpus to be included in the EmoSim508 dataset, annotators have rated them as not related. However, we can also see that the unrelated emoji only cover 29.7% (153/508 for Q1) of the dataset, leaving 70.3% of the dataset with diverse relationships.

4.4 Evaluating Emoji Embedding Models

In this section, we discuss how we evaluated the different emoji embedding models using EmoSim508 as a gold standard dataset. We generated nine ranked lists of emoji pairs based on emoji similarity scores, one ranked list representing the EmoSim508 emoji similarity and eight ranked lists representing each emoji embedding model obtained under different corpus settings. Treating EmoSim508’s emoji similarity ranks as our ground truth emoji rankings, we use Spearman’s rank correlation coefficient⁷ (Spearman’s ρ) to evaluate how well the emoji similarity rankings generated by our emoji embedding models align with the emoji similarity rankings of the gold standard dataset. We used Spearman’s ρ because we noticed that our emoji annotation distribution does not follow a normal distribution. The rank correlation obtained for each setting (multiplied by 100 for display purposes) is shown in Table 4.1. Based on the rank correlation results, we notice that emoji embedding models learned over emoji descriptions (*Sense.Desc.*) moderately correlate ($40.0 < \rho < 59.0$) with the gold standard results, whereas all other models show a strong correlation ($60.0 < \rho < 79.0$). All results reported in Table 4.1 are statistically significant ($p < 0.05$).

We observe that the emoji embeddings learned on sense labels correlate best with the emoji similarity rankings of the gold standard dataset. We further looked into what could be the reason for

⁷<https://goo.gl/ZA4zDP>

Table 4.1: Spearman’s Rank Correlation Results.

Emoji Embedding Model	$\rho \times 100$ for each Corpus	
	Google News	Twitter
<i>(Sense_Desc.)</i>	49.0	46.6
<i>(Sense_Label)</i>	76.0	70.2
<i>(Sense_Def.)</i>	69.5	66.9
<i>(Sense_All)</i>	71.2	67.7

emoji sense labels-based embedding models (*Sense_Label*) to outperform other models. Past work suggests that having access to lengthy emoji sense definitions could improve the performance of the emoji embedding models [Eisner et al. 2016; Pohl et al. 2017]. For the 158 emoji in EmoSim508 dataset, emoji meanings were represented using 25 words on average when using the emoji descriptions; 12 words when using the emoji sense labels; 567 words when using the emoji sense definitions; and 606 words when all above definitions were combined. All our emoji embedding definitions have more words (at least twice as many) than past work [Eisner et al. 2016], but we notice that emoji sense labels are very specific words that only describe emoji meanings, unlike the words in emoji sense descriptions and emoji sense definitions. In contrast, emoji descriptions and emoji sense definitions provide more words describing how an emoji is shown on different platforms or how an emoji should be used in a sentence while describing the emoji’s meaning. These unrelated words in emoji definitions may well be the reason for degraded performance of (*Sense_Desc.*), (*Sense_Def.*) and (*Sense_All*) embeddings. Thus, access to quality sense labels are of vital importance for learning good emoji embeddings.

4.5 Emoji Embeddings at Work

To show that our emoji embedding models can be used in real-world NLP tasks⁸, we set up a sentiment analysis experiment using the gold standard dataset used in [Novak et al. 2015]. We selected this dataset because Barbieri *et al.*'s [Barbieri et al. 2016] and Eisner *et al.*'s [Eisner et al. 2016] models have already been evaluated on this dataset. Thus, it allows us to compare our embedding models with theirs. The gold standard dataset consists of nearly 66,000 English tweets, labelled manually for positive, neutral or negative sentiment. The dataset is divided into a testing set that consist of 51,679 tweets, where 11,700 of them contain emoji, and a training set that consist of 12,920 tweets with 2,295 of them contain emoji. In both the training set and the test set, 46% of tweets are labeled neutral, 29% are labeled positive, and 25% are labeled negative. Thus, the dataset is reasonably balanced.

To represent a training instance in our sentiment analysis dataset, we replaced every word in a tweet using the different embedding models learned for that word by using different text corpora. We also replaced every emoji in the tweet with its representation from a particular emoji embedding model we learned. Table 4.2 shows the results we obtained for the sentiment analysis task when using different emoji embeddings. Here, Google News + (*Sense_Desc.*) means that all words in the tweets in the gold standard dataset are replaced by their corresponding word embedding models learned by the Google News corpus and all emoji are replaced by their corresponding emoji embeddings obtained by the (*Sense_Desc.*) model. We report classification accuracies for: (i) the whole testing dataset ($N = 12,920$); (ii) all tweets with emoji ($N = 2,295$); (iii) 90% of the most frequently used emoji in the test set ($N = 2,186$); and (iv) 10% of the least frequently used emoji in the test set ($N = 308$). We trained a Random Forrest (RF) classifier and a Support Vector Machine (SVM) classifier using each test data segment. We selected those two classifier models as they are commonly used for text classification problems, including the sentiment analysis experiment conducted by Eisner

⁸Please note that our main goal is to demonstrate the usefulness of the learned embedding models and not to develop a state-of-the-art sentiment analysis algorithm.

Table 4.2: Accuracy of the Sentiment Analysis task using Emoji Embeddings.

Word Embedding Model	Classification accuracy on sections of testing dataset							
	N = 12,920		N = 2,295		N = 2,186		N = 308	
	RF	SVM	RF	SVM	RF	SVM	RF	SVM
Google News + emoji2vec	59.5	60.5	54.4	59.2	55.0	59.5	54.5	55.2
Google News + (<i>Sense_Desc.</i>)	58.7	61.9	50.6	55.0	49.7	55.3	45.4	50.0
Twitter + (<i>Sense_Desc.</i>)	60.2	62.5	55.1	56.7	53.8	57.3	53.5	53.2
Google News + (<i>Sense_Label</i>)	60.3	63.3	55.0	61.8	56.8	62.3	54.2	59.0
Twitter + (<i>Sense_Label</i>)	60.7	63.6	57.3	60.8	57.5	61.5	56.1	58.4
Google News + (<i>Sense_Def.</i>)	59.0	62.2	50.3	55.0	51.1	55.2	48.0	50.6
Twitter + (<i>Sense_Def.</i>)	60.0	62.4	53.6	56.2	53.7	56.7	50.6	50.6
Google News + (<i>Sense_All</i>)	59.1	62.2	50.8	55.1	50.2	55.3	50.0	50.6
Twitter + (<i>Sense_All</i>)	60.3	62.4	53.1	57.6	54.1	56.8	54.5	50.0

et al. [Eisner et al. 2016] on the same gold standard dataset. Table 4.2 summarizes the results obtained in the sentiment analysis task. Following Eisner *et al.* [Eisner et al. 2016], we also report the accuracy of the sentiment analysis task, which allows us to compare our embedding models with theirs. Accuracy is measured in settings where the testing dataset is divided into four groups based on the availability of tweets with emoji in each group. We find that the embeddings learned over emoji sense labels perform best in the sentiment analysis task, outperforming the previous best emoji embedding model [Eisner et al. 2016] with an improvement of 7.73%. This embedding model also yielded the best similarity ranking as per Spearman’s Rank Correlation test.

As discussed in Section 4.4, we believe that the inclusion of words that are highly related to emoji meanings make emoji embeddings over sense labels to learn better models to represent the meaning of an emoji, hence, outperform the other models in the sentiment analysis task. We also notice

that Twitter-based emoji embedding models continue to outperform Google News-based embedding models in the majority of the test run settings. Past research on social media text processing suggests that NLP tools designed for social media text processing outperform NLP tools designed for well-formed text processing on the same task [Wijeratne et al. 2017a]. We believe this could be the reason why Twitter-based models continue to outperform Google News-based models. Our results, which continue to outperform Eisner *et al.*'s model [Eisner et al. 2016], prove that the use of emoji descriptions, sense labels, and emoji definitions to model emoji meanings has resulted in learning better emoji embedding models.

4.6 Summary

This chapter presented how the semantic similarity of emoji can be calculated by utilizing the machine-readable emoji sense definitions available in EmojiNet. We looked at how different types of emoji meanings available in EmojiNet can be used to learn emoji embedding models and evaluated the embedding models using EmoSim508 dataset. To show a real-world use-case of the learned emoji embedding models, we used them in a sentiment analysis task and presented the results. EmoSim508 dataset and our emoji embedding models can be downloaded from <http://emojinet.knoesis.org/emosim508.php>. Next, we will explore how EmojiNet can be used to solve emoji sense disambiguation tasks.

5

Applications of EmojiNet – Emoji Sense Disambiguation

5.1 Overview

People use emoji to add color and whimsiness to their messages [Kelly and Watts 2015] and to articulate hard to describe emotions [emo 2015]. Perhaps by design, emoji were defined with no rigid semantics attached to them¹, allowing people to develop their own use and interpretation. Thus, similar to words, emoji can take on different meanings depending on context and part-of-speech [Miller et al. 2016]. For example, consider the three emoji 😂, 🙌, and 🙈 and their use in multiple tweets in Figure 5.1. Depending on context, we see that each of these emoji can take on wildly different meanings. People use the 😂 emoji to mean laughter, sadness, and humor; the 🙌 emoji to express praying, high-fiving or thanking; and the 🙈 emoji to refer to monkeys, hiding or blindness. An application designed to disambiguate emoji senses can be used improve sentiment and emotion analysis applications. For example, consider the emoji 😂, which can take the meanings *happy* and *sad* based on the context in which it has been used. Current sentiment

¹http://www.unicode.org/faq/emoji_dingbats.html#4.0.1




					
Sense	Example	Sense	Example	Sense	Example
Laugh (noun)	Can't stop laughing 😂	Pray (verb)	Pray for my family, god gained an angel today 🙏	Monkey (Noun)	Got a pet monkey 🐒
Crying (verb)	My knee hurts, already in tears 😭😭	Highfive (noun)	We did it man! High-fives all around 🙌🙌🙌	Hiding (verb)	The dog was hiding behind the door 🐕
Hilarious (Adjective)	Central Intelligence was damn hilarious! 😂	Thanks (noun)	Thank you so much for taking care of the baby 🙏	Blind (verb)	I'm blind with no lights on. Can't see anything 🐒

Figure 5.1: Emoji Usage in Social Media with Multiple Senses.

analysis applications do not differentiate among these two meanings when they process 😂.

This chapter introduces the emoji sense disambiguation problem and discusses how EmojiNet can be used as an emoji sense inventory to solve it. There is a lack of availability in emoji sense-tagged data which can be used to solve the emoji sense disambiguation problem in a supervised learning setting. We show that EmojiNet can be utilized as an alternate knowledge-based method to solve the emoji sense disambiguation problem, below.

5.2 Proposed Approach

To show EmojiNet could be used as a sense inventory for emoji sense disambiguation, we first selected 25 emoji which have shown to be interpreted differently when used in communication by previous research [Miller et al. 2016]. Then we take the Twitter corpus that we used to train the word embedding model discussed in Section 3.5 and randomly select 50 tweets for each of the 25 emoji. We select tweets that contain only one emoji anywhere in the middle of the tweet. To disambiguate the sense of an emoji in a tweet, we compare the context of the emoji in the tweet with the contexts of each emoji sense for that emoji obtained from EmojiNet. This tweet context is defined as all words surrounding the emoji. We define three sets of contexts for an emoji sense based on the three different datasets we used to generate them:

- **BabelNet-based context:** This is the set of words coming from BabelNet sense definitions











which we extracted for an emoji. It also includes the words in sense labels and in examples that show how to use those sense labels in sentences,

- **Twitter-based context:** This is the set of context words learned by using the Twitter word embedding model for the emoji from Section 3.5. Each word in the BabelNet-based context is expanded by using the related words learned from the Twitter word embedding model, and
- **News-based context:** This is the set of context words learned by using the Google News word embedding model for an emoji from Section 3.5. Each word in the BabelNet-based context is expanded by using the related words learned from the Google News word embedding model.

To find the sense of an emoji in a tweet, we calculate the context overlap between the context of the emoji in the tweet with the context words taken from each of the above three sets. Following past studies, the sense with the highest context word overlap is assigned to the emoji at the end of a successful run of the algorithm [Vasilescu et al. 2004]. We then asked two human judges to evaluate the emoji senses assigned to the emoji in our tweet dataset. We asked the judges to label the sense assignment as *correct* if they think that the chosen sense for an emoji in a tweet is the most appropriate sense that could be assigned to it from EmojiNet or *incorrect* if they do not think the sense is appropriately assigned for the emoji in a tweet. The agreement between the two judges for this task measured by Cohen’s kappa coefficient was 0.6878, which is considered to be a good agreement.

Table 5.1 lists the top 10 emoji based on their sense disambiguation accuracy. We define the emoji sense disambiguation accuracy for an emoji as the ratio between the number of correctly sense disambiguated messages (tweets) and the number of total sense disambiguated messages for that emoji. Among the 25 emoji in our dataset, 🤔 gives the highest sense disambiguation accuracy of 0.61. We observe that Twitter-based context vectors outperforms the other two context vectors constantly, except for disambiguating the sense of 🍷. This observation aligns with what past research on social media text processing suggest us, which is, tools designed for well-formed text processing

Table 5.1: Top 10 Emoji based on the Emoji Sense Disambiguation Accuracy (in % values).

Emoji	BabelNet-based	Twitter-based	News-based
	24.48	61.22	32.65
	20.93	60.00	59.45
	16.27	56.41	41.46
	12.00	56.00	29.16
	16.66	43.58	52.17
	18.75	51.21	41.17
	15.21	48.57	43.24
	20.45	47.72	13.63
	12.00	46.51	37.50
	27.08	44.18	38.63
Avg. Accuracy	18.38	51.54	38.91

will not work well when used for ill-formatted text processing [Ritter et al. 2011]. The average number of Twitter-based context words for an emoji sense definition was very high compared to that of BabelNet-based contexts. This align with the past research too, that the disambiguation results can be improved when we increase the number of context words in the sense definitions [Vasilescu et al. 2004]. These evaluation results validates the importance of the improvements we made to EmojiNet by introducing context word vectors learned by Twitter and Google News corpora. It also validates the fact that EmojiNet can be successfully used as an emoji knowledgebase that can enable using existing natural language processing techniques for developing algorithms for emoji understanding tasks.

5.3 Summary

This chapter presented the emoji sense disambiguation problem and discussed how EmojiNet can be used as an emoji sense inventory to solve it. It discussed the challenges in solving the emoji sense disambiguation problem using supervised methods and discussed how EmojiNet can be used as a knowledgebase to solve the problem in a knowledge-based setting. It presented emoji sense disambiguation experiments carried out on 25 emoji that are identified in previous research as highly ambiguous. The results showed the potential of EmojiNet as a framework to solve the emoji sense disambiguation problem.

6

Conclusion and Future Work

6.1 Overview

This section concludes the dissertation. In this section, we brief what was covered in the dissertation. Then, we discuss potential future research related to building emoji sense inventories, emoji similarity calculation, and emoji sense disambiguation. We also discuss other future research directions that are important to further emoji research.

6.2 EmojiNet: Building a Machine-Readable Emoji Sense Inventory

We presented the construction of EmojiNet, the first ever machine-readable sense inventory to understand the meanings of emoji. It integrates three different emoji resources from the Web to extract emoji senses and aligns those senses with BabelNet. We discussed the process involved in building EmojiNet and presented how we evaluated those processes. We then explored how EmojiNet can be used to solve multiple emoji understanding tasks. The EmojiNet framework, along with the complementary sense embeddings, vendor-specific emoji meanings, and its REST API with

documentation are publicly released at <http://emojinet.knoesis.org/>.

6.3 Emoji Similarity Calculation

Emoji similarity is shown to be important for many applications such as: (i) emoji-based search [Capallo et al. 2015]; (ii) sentiment analysis [Barbieri et al. 2016; Eisner et al. 2016; Novak et al. 2015], and (iii) mobile keyboard design [Pohl et al. 2017]. However, the notion of the similarity of two emoji is very broad [Wijeratne et al. 2017b]. We presented a method to measure the semantic similarity of emoji, such that the similarity measure reflects the likeness of the emoji meaning, interpretation or intended use. Using the emoji descriptions, emoji sense labels and emoji sense definitions extracted from EmojiNet, on top of two different training corpora obtained from Twitter and Google News, we explored multiple emoji embedding models to measure emoji similarity. We showed that we can generate word vectors that encode emoji meanings by representing the emoji meanings using word embedding models. With the help of ten human annotators who are knowledgeable about emoji, we created EmoSim508 dataset, which consists of 508 emoji pairs and used it as the gold standard to evaluate how well our emoji embedding models perform in an emoji similarity calculation task. To show a real-world use-case of the learned emoji embedding models, we used them in a sentiment analysis task and presented the results. We released the EmoSim508 dataset at <http://emojinet.knoesis.org/emosim508.php>.

6.4 Emoji Sense Disambiguation

The state-of-the-art emoji processing applications do not account for the contextual differences in emoji meaning and interpretation, which poses challenges for applications that utilize emoji. Thus, similar to the word sense disambiguation task in natural language processing, applications need to disambiguate the meaning or ‘sense’ of an emoji. We examined this problem and tried to solve it by combining EmojiNet sense definitions with traditional NLP techniques. We demonstrated an

approach that automatically disambiguates the meanings of polysemous emoji. We evaluated our approach using a manually annotated Twitter dataset consists of at least one emoji in each tweet and its meaning in the message context. We showed that EmojiNet sense definitions can be coupled with traditional NLP algorithms to solve the emoji sense disambiguation problem.

6.5 Future Work

The work presented in this dissertation can be further extended in three important directions, namely, (i) building emoji sense dictionaries, (ii) emoji similarity calculation, and (iii) emoji sense disambiguation. We will look at each of these directions in more detail, below.

6.5.1 Building a Machine-Readable Emoji Sense Inventory

The approach we followed to build EmojiNet can be further extended in several ways. One can extend the EmojiNet sense inventory by adding machine processable emoji meanings such as slang terms that were not present in BabelNet but listed as intended meanings by the Unicode Consortium. For example, we noticed that slang terms such as **OMG** are filtered in the data filtration process as **OMG** is not listed as a word in BabelNet. However, it is listed as a concept in BabelNet, which is not associated with a PoS tag, thus, it gets filtered in the process. Therefore, we believe an additional step is needed to identify slang terms which can be very important for certain emoji (e.g., **OMG** is one of the most prominent meanings for ‘Face Screaming in Fear’ 🤪 emoji.). EmojiNet framework can also benefit from a semi-automatic update process to keep EmojiNet up-to-date as and when new emoji are supported by the Unicode Consortium. We noticed that the Unicode Consortium requires new emoji proposals to list the possible uses of those emoji. These emoji usages reflect different emoji meanings that the designers of the emoji have assigned them. Thus, building methods to automatically extract emoji usages from emoji proposals available in the Unicode Consortium Website can be helpful to further extend EmojiNet. EmojiNet currently supports platform-specific

emoji meanings for 40 emoji that are highly misunderstood across mobile platforms. However, there can be other emoji that are mis-classified due to their depictions across platforms, that are not covered in our study. Further studies on platform-specific emoji meanings can contribute to finding these missing emoji meanings in EmojiNet. One could also work on introducing more sophisticated algorithms to solve image processing and word sense disambiguation algorithms we used in evaluating the creation process of EmojiNet.

6.5.2 Emoji Similarity Calculation

Our emoji similarity calculation approach can be further extended as follows. The emoji embedding models can be further extended to understand the differences in emoji interpretations due to how they appear across different platforms or devices. EmojiNet currently lists platform-specific meanings for 40 emoji and those platform-specific meanings were not treated differently when we trained our emoji embedding models. However, once EmojiNet started to carry platform-specific emoji meanings for all emoji standardized by the Unicode Consortium (thus, available in EmojiNet), one could also train platform-specific emoji embeddings following the approaches we discussed in this dissertation. One could also apply our emoji embedding models to other emoji analysis tasks such as emoji-based search. Emoji-based search is now getting the attention of the Internet-based companies. Especially, Internet giants such as YouTubeMusic has already started supporting emoji-based search in their mobile application. It will be interesting to explore whether emoji similarity results could be used to improve the recall in emoji-based search applications. If the gold standard datasets can be created, the emoji similarity methods discussed in this dissertation can be evaluated for improvements in recall in emoji-based search.

6.5.3 Emoji Sense Disambiguation

Our emoji sense disambiguation methods can be further extended by introducing more sophisticated algorithms to solve emoji sense disambiguation problem. For example, Iacobacci *et al.* recently

studied how word embedding-based methods can be used to improve word sense disambiguation tasks [Iacobacci et al. 2016; Bordea et al. 2016]. They experimented with several word vector summation methods including (i) vector concatenation, (ii) vector average, (iii) fractional decay, and (iv) exponential decay [Iacobacci et al. 2016; Bordea et al. 2016]. They also studied different word embedding generation methods including (i) Word2Vec [Mikolov et al. 2013], (ii) Collobert and Weston’s method [Collobert and Weston 2008], and (iii) retrofitting of word embeddings [Faruqui et al. 2015]. One can incorporate the findings of Iacobacci *et al.* [Iacobacci et al. 2016] to introduce state-of-the-art word embedding models to solve emoji sense disambiguation problem. Another possible direction for future research would be creating evaluation datasets for emoji sense disambiguation. For example, evaluation datasets for emoji sense disambiguation experiments we carried out as part of this dissertation were consist of randomly selected tweets with emoji. However, these datasets can be further improved by hand-selecting example emoji usages where emoji are used to replace words in Tweets. We didn’t consider platform-specific emoji meanings when conducting our emoji sense disambiguation experiments. Specifically, social media platforms such as Twitter records the platform that a tweet generated from and disseminates it via their programming API. Thus, one could use the platform details available in Twitter (via the `source` field in Twitter JSON object [Wijeratne et al. 2017]) to further improve the emoji sense disambiguation algorithms.

6.5.4 Emoji Prediction

Emoji prediction is the problem of predicting an emoji to a given text segment. In the simplest form, it tries to predict an emoji at the end of a sentence (or a tweet) in a way that the predicted emoji captures what is conveyed in the sentence. Several recent research studies have tried to solve the emoji prediction problem in a supervised setting [Barbieri et al. 2017; Barbieri et al. 2018; Barbieri et al. 2018] and have achieved prediction accuracies that range from 13% to 48%. This tells us that the emoji prediction is a challenging problem. Past research has shown that supervised learning approaches can be improved by using carefully generated domain knowledge [Sheth et al. 2017].

Thus, it will be interesting to explore the possibility of utilizing the emoji senses knowledge available in EmojiNet to improve the accuracy of emoji prediction problem. Working on the novel methods to incorporate EmojiNet knowledgebase into deep learning architectures so that the resulting models can be used to solve the emoji prediction problem will be a challenging task.

6.6 Going Forward

In this dissertation, we looked at how external knowledge on emoji meanings can be used to improve emoji understanding tasks. We also looked at ways to incorporate traditional NLP algorithms into our solutions. Emoji usage is complex in nature. Compared to regular languages, emoji do not have a clear hierarchical grammar [McCulloch and Gawne 2018]. As suggested by Danesi in [Danesi 2016], we also believe that emoji would continue to exist as a complementary writing element for alphabetic writing systems. Therefore, we believe that efforts to use existing NLP systems with emoji-specific language processing systems will be crucial for the further development of emoji research. We believe having access to emoji datasets will also play a major role in furthering emoji research. For example, we created ground truth datasets for emoji similarity and emoji sense disambiguation. We believe these datasets can further be improved. For example, as discussed in Section 6.5.3, we believe emoji sense disambiguation datasets should include more hand-selected instances. Specifically, we should include training instances where emoji replaces words. So far, many researchers have ignored the role that the emoji play in a text message when analyzing them. For example, emoji sense disambiguation and emoji prediction approaches discussed in this dissertation do not consider the position of an emoji in a text message or the role of the emoji (e.g., tone modification, replace words, emphasize an existing word etc.) when addressing those emoji understanding tasks. However, we believe having an understanding of where an emoji appears in a text and the role it plays could further improve the algorithms we develop. Therefore, further research should shed a light on investigating such problems to measure their effects on the performance of the emoji understanding tasks.

Platform-specific emoji meanings are still common for many emoji. We have recently seen major mobile platform providers (such as Apple and Android) trying to minimize the differences in their emoji designs and trying to agree on universal emoji designs. For example, in 2018, all major mobile and web platform providers decided to replace the gun emoji with a water pistol emoji, following Apple, who first changed their gun emoji to a water pistol in 2016¹. However, it is well known that mobile platform providers like to maintain their own versions of emoji pictographs, except for rare instances like the gun emoji. Thus, we expect emoji miscommunication due to platform-specific emoji depictions to continue to exist. We believe platform-specific emoji meanings can be valuable to address emoji miscommunications. However, there hasn't been scalable efforts to learn platform-specific emoji representations, mainly due to the challenges in collecting platform-specific emoji datasets. Even though emoji data which are generated from each mobile or web platform is hard to find, we can extend EmojiNet to hold platform-specific emoji meanings, which can then be used to learn platform-specific emoji embeddings by using the emoji embedding learning methods discussed in Chapter 4. We believe that emoji sense disambiguation and emoji prediction will continue to be hard-to-solve emoji understanding tasks. We are hopeful that EmojiNet would continue be a valuable resource to address those open problems.

6.7 Summary

This section briefly presented what was covered in the dissertation. It also discussed potential future research related to building emoji sense inventories, emoji similarity calculation, emoji sense disambiguation, and emoji prediction. Finally, it discussed the challenges in furthering emoji research.

¹<https://www.abc.net.au/news/2018-04-29/gun-emoji-replaced-with-toy-water-pistol-across-platforms/>

Appendices






























A

Platform-specific Emoji in EmojiNet

A.1 Overview

Here, we list all the platform-specific emoji in EmojiNet. We've collected platform-specific emoji meanings for the 40 emoji listed in Table A.1 using an Amazon Mechanical Turk experiment and the results can be downloaded from <http://emojinet.knoesis.org/dataset/emojiplatformresults.zip>. For the illustration purposes, we've only shown Twitter platform-specific emoji images in Table A.1. All platform-specific emoji images we considered in this study can be found at http://emojinet.knoesis.org/dataset/vendorspecific_emoji.html.

Table A.1: Platform-specific Emoji in EmojiNet.

Unicode	Emoji (Twitter)	Unicode	Emoji (Twitter)
U+263A		U+1F301	
U+1F375		U+1F41E	
U+1F428		U+1F443	
U+1F46F		U+1F473	
U+1F47B		U+1F483	
U+1F486		U+1F4BB	
U+1F4F1		U+1F52B	
U+1F575		U+1F601	
U+1F602		U+1F604	
U+1F605		U+1F606	
U+1F60A		U+1F60B	
U+1F60C		U+1F60D	
U+1F60E		U+1F60F	
U+1F612		U+1F614	
U+1F618		U+1F622	
U+1F629		U+1F62A	
U+1F62D		U+1F631	
U+1F647		U+1F648	
U+1F64C		U+1F64F	
U+1F923		U+1F933	

B

EmojiNet REST API Calls

B.1 Overview

This chapter explains EmojiNet REST API calls. It lists each REST API call with corresponding input parameters and sample requests. Currently, EmojiNet REST API supports 7 methods that can be used to retrieve information about emoji stored in EmojiNet. They are: (i) Get Emoji Information, (ii) Get Emoji Images, (iii) Get Noun Meanings for Emoji, (iv) Get Verb Meanings for Emoji, (v) Get Adjective Meanings for Emoji, (vi) Get Twitter Word Embeddings for Emoji, and (vii) Get Google News Word Embeddings for Emoji. Each method in the EmojiNet REST API is discussed below by providing the functionality of the method, method signature (URL field), URL parameters of the method, and a sample request that shows how the method is invoked. For a detailed description of each method signature including sample responses and error codes, please refer to the Online EmojiNet REST API available at - <http://emojinet.knoesis.org/api.php>.

Table B.1: Get Emoji Information.

Function	Get Emoji Information
Description	This method returns information about a given emoji including its Unicode, name, shortcode, description, keywords, category, and related emoji.
URL	<code>/emoji/:emojiunicode</code>
URL Parameters	<code>:emojiunicode=[unicodeString]</code>
Sample Request	<code>http://emojinet.knoesis.org/api/emoji/U0001F64C</code>

Table B.2: Get Emoji Images.

Function	Get Emoji Images
Description	This method returns information about vendor-specific images stored in EmojiNet for a given emoji. A binary representation of the image will be returned with the corresponding vendor name.
URL	<code>/emoji/images/:emojiunicode</code>
URL Parameters	<code>:emojiunicode=[unicodeString]</code>
Sample Request	<code>http://emojinet.knoesis.org/api/emoji/images/U0001F64C</code>

Table B.3: Get Noun Meanings for Emoji.

Function	Get Noun Meanings for Emoji
Description	This method returns information about all noun meanings stored in EmojiNet for a given emoji. It will return NULL if there are no noun meanings available for the emoji.
URL	<code>/emoji/noun/:emojiunicode</code>
URL Parameters	<code>:emojiunicode=[unicodeString]</code>
Sample Request	<code>http://emojinet.knoesis.org/api/emoji/noun/U0001F64C</code>

Table B.4: Get Verb Meanings for Emoji.

Function	Get Verb Meanings for Emoji
Description	This method returns information about all verb meanings stored in EmojiNet for a given emoji. It will return NULL if there are no verb meanings available for the emoji.
URL	<code>/emoji/verb/:emojiunicode</code>
URL Parameters	<code>:emojiunicode=[unicodeString]</code>
Sample Request	<code>http://emojinet.knoesis.org/api/emoji/verb/U0001F64C</code>

Table B.5: Get Adjectives Meanings for Emoji.

Function	Get Adjectives Meanings for Emoji
Description	This method returns information about all adjective meanings stored in EmojiNet for a given emoji. It will return NULL if there are no adjective meanings available for the emoji.
URL	/emoji/adjective/:emojiunicode
URL Parameters	:emojiunicode=[unicodeString]
Sample Request	http://emojinet.knoesis.org/api/emoji/adjective/U0001F602

Table B.6: Get Twitter Word Embeddings for Emoji.

Function	Get Twitter Word Embeddings for Emoji
Description	This method returns Twitter-based context words learned for a given emoji sense. The method expects a BabelNet sense ID as the input and returns a list of words learned by our Twitter-based word embedding model for the given sense definition.
URL	/sensevec/twitter/:babelnetsenseID
URL Parameters	:babelnetsenseID=[babelnetsenseIDString]
Sample Request	http://emojinet.knoesis.org/api/sensevec/twitter/bn:00104206a

Table B.7: Get Google News Word Embeddings for Emoji.Emoji (Twitter)

Function	Get Google News Word Embeddings for Emoji
Description	This method returns Google News-based context words learned for a given emoji sense. The method expects a BabelNet sense ID as the input and returns a list of words learned by our Google News-based word embedding model for the given sense definition.
URL	/sensevec/google/:babelnetsenseID
URL Parameters	:babelnetsenseID=[babelnetsenseIDString]
Sample Request	http://emojinet.knoesis.org/api/sensevec/google/bn:00104206a

References

2015. Emogi research team - 2015 emoji report.
- AI, W., LU, X., LIU, X., WANG, N., HUANG, G., AND MEI, Q. 2017. Untangling emoji popularity through semantic embeddings. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. Montreal, Canada, 2–11.
- ARTSTEIN, R. AND POESIO, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4, 555–596.
- BAI, J. 2018. Emoji & communication. *The Ling Thing* 8, 18.
- BALASURIYA, L., WIJERATNE, S., DORAN, D., AND SHETH, A. 2016. Finding street gang members on twitter. In *Proceedings of The 2016 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*. Vol. 8. San Francisco, CA, USA, 685–692.
- BARBER, M. J. AND CLARK, J. W. 2009. Detecting network communities by propagating labels under constraints. *Physical Review E* 80, 2, 026129.
- BARBIERI, F., BALLESTEROS, M., RONZANO, F., AND SAGGION, H. 2018. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (HLT-NAACL)*. Vol. 2. 679–686.

- BARBIERI, F., BALLESTEROS, M., AND SAGGION, H. 2017. Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 2, Short Papers*. Vol. 2. 105–111.
- BARBIERI, F., CAMACHO-COLLADOS, J., RONZANO, F., ANKE, L. E., BALLESTEROS, M., BASILE, V., PATTI, V., AND SAGGION, H. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation (Sem-Eval)*. 24–33.
- BARBIERI, F., ESPINOSA-ANKE, L., BALLESTEROS, M., SAGGION, H., ET AL. 2017. Towards the understanding of gaming audiences by modeling twitch emotes. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (W-NUT 2017); 2017 Sep 7; Copenhagen, Denmark. Stroudsburg (PA): ACL; 2017. p. 11-20*. ACL (Association for Computational Linguistics).
- BARBIERI, F., KRUSZEWSKI, G., RONZANO, F., AND SAGGION, H. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM International Conference on Multimedia. MM '16*. ACM, New York, NY, USA, 531–535.
- BARBIERI, F., RONZANO, F., AND SAGGION, H. 2016. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia.
- BASILE, P., CAPUTO, A., AND SEMERARO, G. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. 1591–1600.
- BORDEA, G., LEFEVER, E., AND BUITELAAR, P. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1081–1091.
- CAMACHO-COLLADOS, J., PILEHVAR, M. T., AND NAVIGLI, R. 2015. Nasari: a novel approach to a

- semantically-aware representation of items. In *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HTL-NAACL)*. 567–577.
- CAPPALLO, S., MENSINK, T., AND SNOEK, C. G. 2015. Query-by-emoji video search. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. 735–736.
- CHANDLER, D. 2007. *Semiotics: the basics*. Routledge.
- COLLOBERT, R. AND WESTON, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*. ACM, 160–167.
- CRAMER, H., DE JUAN, P., AND TETREAULT, J. 2016. Sender-intended functions of emojis in us messaging. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. MobileHCI '16*. ACM, New York, NY, USA, 504–509.
- DANESI, M. 2016. *The semiotics of emoji: The rise of visual language in the age of the internet*. Bloomsbury Publishing.
- DIMSON, T. 2015. Emojineering part 1: Machine learning for emoji trends. *Instagram Engineering Blog*.
- DONATO, G. AND PAGGIO, P. 2017. Investigating redundancy in emoji use: Study on a twitter based corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 118–126.
- EISNER, B., ROCKTÄSCHEL, T., AUGENSTEIN, I., BOŠNJAK, M., AND RIEDEL, S. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media at EMNLP 2016 (SocialNLP at EMNLP 2016)*. Austin, Texas, USA.

- EVANS, V. 2017. *The emoji code: The linguistics behind smiley faces and scaredy cats*. Picador USA.
- FARUQUI, M., DODGE, J., JAUHAR, S. K., DYER, C., HOVY, E., AND SMITH, N. A. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. 1606–1615.
- FELBO, B., MISLOVE, A., SØGAARD, A., RAHWAN, I., AND LEHMANN, S. 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 (EMNLP)*. 1615–1625.
- GONZÁLEZ-IBÁNEZ, R., MURESAN, S., AND WACHOLDER, N. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 581–586.
- GUHA, R., MCCOOL, R., AND MILLER, E. 2003. Semantic search. In *Proceedings of the 12th International Conference on World Wide Web (WWW)*. 700–709.
- HARRIS, Z. S. 1954. Distributional structure. *Word* 10, 2-3, 146–162.
- HAYES, A. F. AND KRIPPENDORFF, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*.
- HERRING, S. AND DAINAS, A. 2017. Nice picture comment! graphicons in facebook comment threads. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*. Waikoloa, Hawaii, USA, 2185–2194.
- HERRING, S. C. AND DAINAS, A. R. 2018. Receiver interpretations of emoji functions: A gender perspective. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018)*. Stanford, CA.

- HILL, F., REICHART, R., AND KORHONEN, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. In *Computational Linguistics*.
- HU, T., GUO, H., SUN, H., NGUYEN, T. T., AND LUO, J. 2017. Spice up your chat: The intentions and sentiment effects of using emojis. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. 102–111.
- HU, Y., TALAMADUPULA, K., AND KAMBHAMPATI, S. 2013. Dude, srsly?: The surprisingly formal nature of twitter’s language. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM)*. Boston, USA, 244–253.
- HUANG, E. H., SOCHER, R., MANNING, C. D., AND NG, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Meeting of the Association for Computational Linguistics (ACL)*. 873–882.
- IACOBACCI, I., PILEHVAR, M. T., AND NAVIGLI, R. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1. 897–907.
- ILLENDULA, A. AND YEDULLA, M. R. 2018. Learning emoji embeddings using emoji co-occurrence network graph. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018)*. Stanford, CA.
- JIANG, J. A., BRUBAKER, J. R., AND FIESLER, C. 2017. Understanding diverse interpretations of animated gifs. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1726–1732.
- JOSHI, A., BHATTACHARYYA, P., AND CARMAN, M. J. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)* 50, 5, 73.
- KELLY, R. AND WATTS, L. 2015. Characterising the inventive appropriation of emoji as relation-

- ally meaningful in mediated close personal relationships. In *Proceedings of the 14th European Conference on Computer Supported Cooperative Work (ECSCW)*. Oslo, Norway.
- KENTER, T., BORISOV, A., AND DE RIJKE, M. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (ACL)*. 941–951.
- LEVY, O., GOLDBERG, Y., AND DAGAN, I. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3, 211–225.
- LIKERT, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.
- LOCKE, J. 1841. *An essay concerning human understanding*.
- MAATEN, L. V. D. AND HINTON, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9, Nov, 2579–2605.
- MCCULLOCH, G. AND GAWNE, L. 2018. Emoji grammar as beat gestures. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018)*. Stanford, CA.
- MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. 2013. Efficient estimation of word representations in vector space. *CoRR*.
- MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G., AND DEAN, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13*. Curran Associates Inc., USA, 3111–3119.
- MIKOLOV, T., YIH, W.-T., AND ZWEIG, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the Human Language Technologies: The 2013 Annual Conference*

- of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*. Vol. 13. 746–751.
- MILLER, H., THEBAULT-SPIEKER, J., CHANG, S., JOHNSON, I., TERVEEN, L., AND HECHT, B. 2016. Blissfully happy or ready to fight: Varying interpretations of emoji. In *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*. Cologne, Germany, 259–268.
- MILLER, H. J., KLUVER, D., THEBAULT-SPIEKER, J., TERVEEN, L. G., AND HECHT, B. J. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*. Montreal, Canada, 152–161.
- MORO, A., NAVIGLI, R., TUCCI, F. M., AND PASSONNEAU, R. J. 2014. Annotating the masc corpus with babelnet. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland, 4214–4219.
- NA’AMAN, N., PROVENZA, H., AND MONTOYA, O. 2017. Varying linguistic purposes of emoji in (twitter) context. In *Proceedings of the 55th Association of Computational Linguistics (ACL), Student Research Workshop*. 136–141.
- NAVIGLI, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41, 2, 10.
- NAVIGLI, R. AND PONZETTO, S. P. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (ACL)*. ACL, 216–225.
- NOVAK, P. K., SMAILOVIĆ, J., SLUBAN, B., AND MOZETIČ, I. 2015. Sentiment of emojis. *PloS one* 10, 12, 1–22.
- PAVALANATHAN, U. AND EISENSTEIN, J. 2015. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*.

- PAVALANATHAN, U. AND EISENSTEIN, J. 2016. More emojis, less:) the competition for paralinguistic function in microblog writing. *First Monday* 21, 11.
- POHL, H., DOMIN, C., AND ROHS, M. 2017. Beyond just text: Semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1, 6.
- REVILLA, M. A., SARIS, W. E., AND KROSNICK, J. A. 2014. Choosing the number of categories in agree–disagree scales. In *Sociological Methods & Research*.
- RITTER, A., CLARK, S., ETZIONI, O., ET AL. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1524–1534.
- SANTHANAM, S., SRINIVASAN, V., GLASS, S., AND SHAIKH, S. 2018. I stand with you: Using emojis to study solidarity in crisis events. In *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018)*. Stanford, CA.
- SANTOS, R. 2010. Java image processing cookbook.
- SHETH, A., PERERA, S., WIJERATNE, S., AND THIRUNARAYAN, K. 2017. Knowledge will propel machine understanding of content: extrapolating from current examples. In *Proceedings of the International Conference on Web Intelligence*. ACM, 1–9.
- SNOW, R., O’CONNOR, B., JURAFSKY, D., AND NG, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 254–263.
- SWIFTKEY, P. 2015. Most-used emoji revealed: Americans love skulls, brazilians love cats, the french love hearts [blog].

- TIGWELL, G. W. AND FLATLA, D. R. 2016. Oh that’s what you meant!: reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI ’16. ACM, 859–866.
- VASILESCU, F., LANGLAIS, P., AND LAPALME, G. 2004. Evaluating variants of the lesk approach for disambiguating words. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.
- WANG, W. 2015. Automatic emotion identification from text. Ph.D. thesis, Wright State University.
- WANG, W., CHEN, L., THIRUNARAYAN, K., AND SHETH, A. P. 2012. Harnessing twitter ”big data” for automatic emotion identification. In *Proceedings of 2012 International Conference on Social Computing (SocialCom) and 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT)*. IEEE, 587–592.
- WEAVER, W. 1955. Translation. *Machine translation of languages 14*.
- WIJERATNE, S., BALASURIYA, L., DORAN, D., AND SHETH, A. 2016. Word embeddings to enhance twitter gang member profile identification. In *Proceedings of the IJCAI Workshop on Semantic Machine Learning (SML 2016)*. CEUR-WS, New York City, NY, 18–24.
- WIJERATNE, S., BALASURIYA, L., SHETH, A., AND DORAN, D. 2016. Emojinet: Building a machine readable sense inventory for emoji. In *Proceedings of the 8th International Conference on Social Informatics (SocInfo 2016)*. Springer International Publishing, Bellevue, WA, USA, 527–541.
- WIJERATNE, S., BALASURIYA, L., SHETH, A., AND DORAN, D. 2017a. Emojinet: An open service and api for emoji sense discovery. In *11th International AAAI Conference on Web and Social Media (ICWSM)*. Montreal, Canada, 437–446.
- WIJERATNE, S., BALASURIYA, L., SHETH, A., AND DORAN, D. 2017b. A semantics-based measure of emoji similarity. In *Proceedings of the International Conference on Web Intelligence*. ACM, 646–653.

- WIJERATNE, S., SHETH, A., BHATT, S., BALASURIYA, L., AL-OLIMAT, H. S., GAUR, M., YAZDAVAR, A. H., AND THIRUNARAYAN, K. 2017. Feature engineering for twitter-based applications. Chapman and Hall. Data Mining and Knowledge Discovery Series, 359–393.
- WOOD, I. AND RUDER, S. 2016. Emoji as emotion tags for tweets. In *Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia*. 76–79.