GAUGING ONLINE AND OFFLINE PUBLIC OPINION FOR SOCIAL MEDIA MONITORING

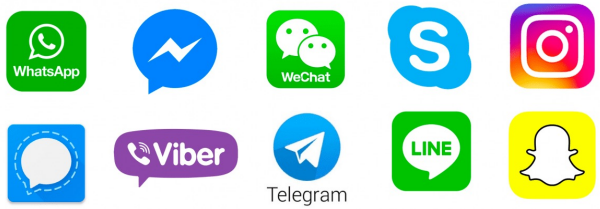
NAME: Ranjan Satapathy, Ph.D. candidate

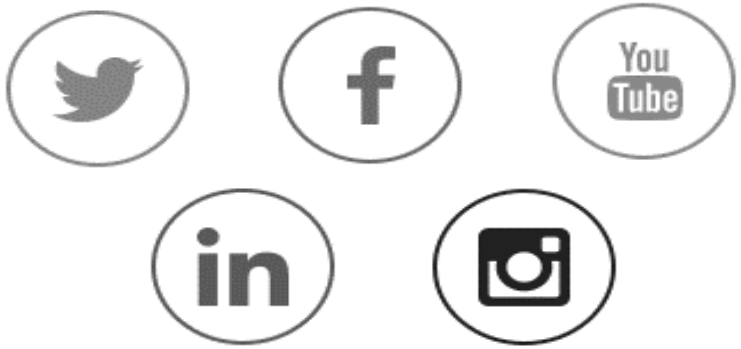
SUPERVISOR: Assoc. Prof. Erik Cambria

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

OBJECTIVE

- The **rise of social media** has led to the development of a new form of texting style known as "**microtext**" [1,2].
  - The pervasiveness of microtext can now be seen in **offline mediums** as well.
  - For example: "I luv this trip to #Singapura".
- Understanding social media content needs analysis of microtext.
- Objective of my work is to build a system to normalize the **out-of-vocabulary (OOV)** words to their **in-vocabulary (IV)** counterparts.
- This will enable the correct interpretation of web content and its application to different NLP tasks.





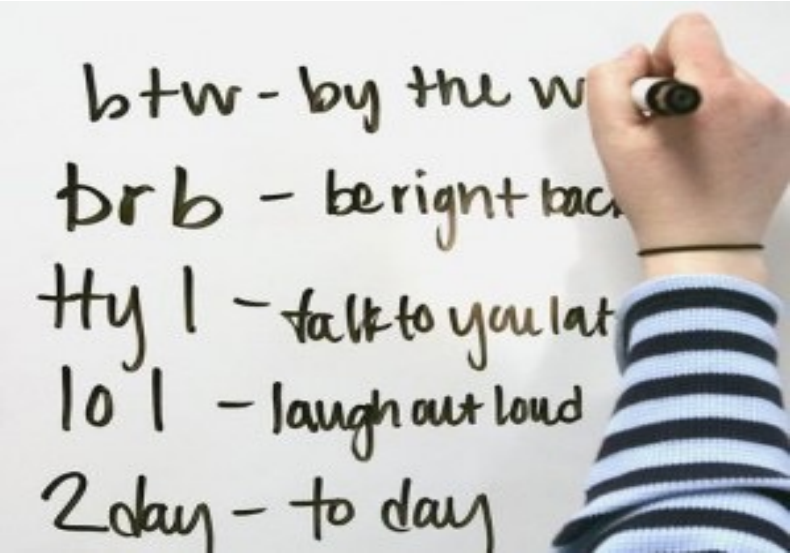
[1] Ellen, J. 2011. All about microtext: A working definition and a survey of current microtext research within artificial intelligence and natural language processing. In Proc. of the Third International Conference on Agents and Artificial Intelligence.
[2] N. . Desai and M. Narvekar, "Normalization of noisy text data," Procedia Computer Science, vol. 45, pp. 127 – 132, 2015, international Conference on Advanced Computing Technologies and Applications (ICACTA).

OUTLINE

- **Introduction**
  - Features of Microtext
  - Classes of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
- **Major Contribution**
- **Conclusion and Future work**

INTRODUCTION:
FEATURES OF MICROTEXT

1. It is **concise**, typically one sentence with shortened words, and sometimes as little as a single word (abbreviations like "hru" for "how are you")
2. it is **written informally** and unedited for quality and thus may use loose grammar, a conversational tone, vocabulary errors, and uncommon abbreviations and acronyms [1]



[1] Xue, D. Yin, and B. D. Davison, "Normalizing Microtext," Analyzing Microtext, pp. 74–79, 2011.

CLASSES OF MICROTEXT

1. CLIPPING
   - Initial
   - Middle
   - Final
2. ACRONYM
3. PHONETIC
4. HYBRID
5. OTHERS

em ***them***

abt ***about***

ack ***acknowledge***

TTYL ***talk to you later***

lyk ***like***

2morro tomorrow

zoh my god ***oh my god***

CLASSES OF MICROTEXT

1. CLIPPING
   - Initial
   - Middle
   - Final
2. ACRONYM
3. PHONETIC
4. HYBRID
5. OTHERS

em ***them***

abt ***about***

ack ***acknowledge***

TTYL ***talk to you later***

lyk ***like***

2morro tomorrow

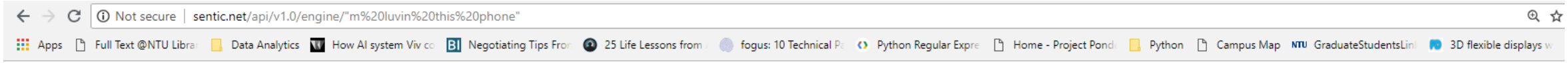
zoh my god ***oh my god***

MOTIVATION: EFFECT OF MICROTEXT ON NLP TASKS
NAMED ENTITY RECOGNITION

- The effectiveness of the **Stanford's named entity recognizer** algorithm falls from 90.8 % to 45.8 % when it is applied to a corpus of Tweets [1].

[1] Liu F, Weng F, Jiang X. A broad-coverage normalization system for social media language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers; 2012. Vol. 1. p. 1035–1044.

MOTIVATION:
POLARITY DETECTION




```
{
  "polarity": {
    "intensity": 0.999999999016844,
    "sentiment": "negative"
  }
}
```


Polarity detection [1] for a text containing microtext (m luvin this phone)

[1] Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018, April). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

MOTIVATION:
SUBJECTIVITY DETECTION

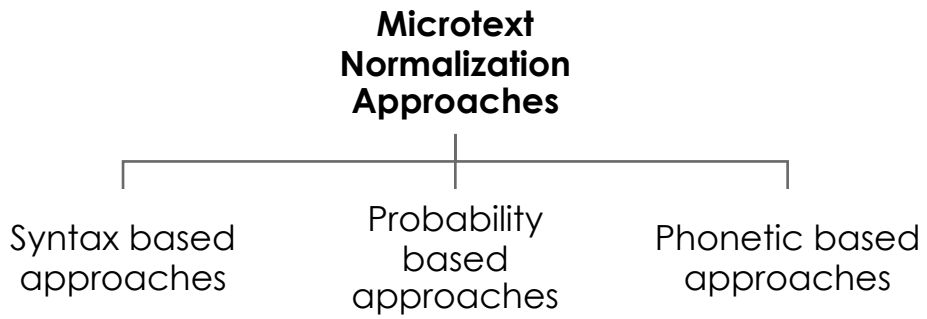
- We built a subjective detection model and trained it on MPQA dataset [1].
- To see the effect of microtext on this proposed model, we crawled twitter and blogs to collect tweets and texts containing microtext.
  - For example:
    - Subjective : "I luv this #nuclearenergy concept, a green initiative" or "i dnt lyk #silkroad connecting whole world".
    - Objective : "Government is trying to start a nuclear power plant soon".
- The **text containing microtext** brought down the F1-measure from **95% to 60%.**

[1] Satapathy, R., Chaturvedi, I., Cambria, E., Ho, S. S., & Na, J. C. (2017). Subjectivity detection in nuclear energy tweets. Computación y Sistemas, 21(4), 657-664.

OUTLINE

- **Introduction**
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
- **Major Contribution**
- **Conclusion and Future work**

LITERATURE SURVEY



- **Syntax based approach** where syntactic units like characters, words or concepts are taken into account to normalize.
- **Probability based approach** where probabilistic models are taken into account to normalize microtext.
- **Phonetic based approach** where phonemic units are taken into account to normalize microtext

LITERATURE SURVEY

Syntax based approaches	Probability based approaches	Phonetic based approaches
• The authors evaluated the proposed model on Tweets and SMS at the word level. MoNoise [1] uses dictionary, Aspell and word embeddings to generate candidates. • Authors in [2] find the most probable candidate from database for a particular OOV after applying Levenshtein distance.	• The authors in [3] participated in W-NUT Lexical Normalization for English Tweets challenge • In [4], the authors propose a hybrid method for multi-class sentiment analysis of micro-blogs, which combines the model and lexicon-based approach.	• Authors in [5] propose a phonetic tree based microtext normalization on English Wiktionary. • In [6], the authors propose a cognitively-inspired normalization technique which is a combination of enhanced letter transformation, visual priming, and string/phonetic similarity.

[1]van der Goot, "MoNoise: A multi-lingual and easy-to-use lexical normalization tool," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 201–206.
[2] N. Desai and M. Narvekar, "Normalization of noisy text data," Procedia Computer Science, vol. 45, pp. 127 – 132, 2015, International Conference on Advanced Computing Technologies and Applications (ICACTA).
[3] S. Leeman-Munk, J. Lester, and J. Cox, "Ncsu sas sam: Deep encoding and reconstruction for normalization of noisy text," in Proceedings of the Workshop on Noisy User-generated Text, 2015, pp. 154–161
[4] S. Yuan, J. Wu, L. Wang, and Q. Wang, "A hybrid method for multi-class sentiment analysis of micro-blogs," in Service Systems and Service Management (ICSSSM), 2016 13th International Conference on. IEEE, 2016, pp. 1–6
[5] R. Khoury, "Phonetic normalization of microtext," in Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on. IEEE, 2015, pp. 1600–1601.
[6] F. Liu, F. Weng, and X. Jiang, "A Broad-Coverage Normalization System for Social Media Language," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, no. July, 2012, pp. 1035–1044.

LIMITATIONS

- Context plays an important role in microtext normalization.
- To build a lexicon based method, it is very difficult to include all the variations.
- No extensive study to show application of microtext normalization on NLP models.

OUTLINE

- **Introduction**
  - Microtext Processing
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
- **Major Contribution**
- **Conclusion and Future work**

UNSUPERVISED (HYBRID) MODEL

- We prepared a polarised microtext lexicon (word-based). We built a framework with a microtext lexicon and a rule-based method for microtext normalization.
- Core of our lexicon is formed from pre-existing websites [1,2,3].
- We added more OOV terms which were frequent and not available in any of the resources.

[1] Internet Slang words - Internet Dictionary - InternetSlang.com
[2] Acronyms and Slang - Simply find the meaning of Acronyms, Internet Slang and Abbreviations online!
[3] NetLingo The Internet Dictionary

SAMPLE LEXICON

OOV	CLASS	POLARITY	IV FORM
u up	PHON	Neutral	(are) you up
dafuq	PHON	Negative	(what) the fuck
AFDA	ACR	Neutral	A Few Days Ago
AFPOE	ACR	Positive	A Fresh Pairs Of Eyes
AFAGAY	ACR	Positive	A Friend As Good As You
ALT	ACR	Positive	A Lot of Talents
catfish	OTHER	Negative	a person assuming a false identity
creeper	OTHER	Negative	a socially invasive person
abt	CLP (M)	Neutral	about
AFT	ACR	Negative	About Fucking Time
AAR	ACR	Neutral	At Any Rate

RULES

Positional knowledge of the OOV	"I love u 2" and "I love my 2 daughters"
Handling emoticons with numerical digit	"m so much in love with my new iphone <3"
Handling repetitions in the text	"that is soooo cooooool"

SOUNDEX ALGORITHM

Soundex hash value is calculated by using the first letter of a name and converting its consonants to digits through a simple lookup table. The pseudo-code of the algorithm is as follows:

1. Retain the first letter of the word;
2. Change all occurrences of the following letters to '0' (zero): 'A', 'E', 'I', 'O', 'U', 'H', 'W', 'Y';
3. Change letters to digits as follows:
   - B, F, P, V → 1
   - C, G, J, K, Q, S, X, Z → 2
   - D,T → 3
   - L → 4
   - M, N → 5
   - R → 6
4. Remove all pairs of consecutive digits;
5. Remove all zeros from the resulting string;
6. Pad the resulting string with trailing zeros and return the first four positions, which will be of the form:
   <uppercase letter> <digit> <digit> <digit>

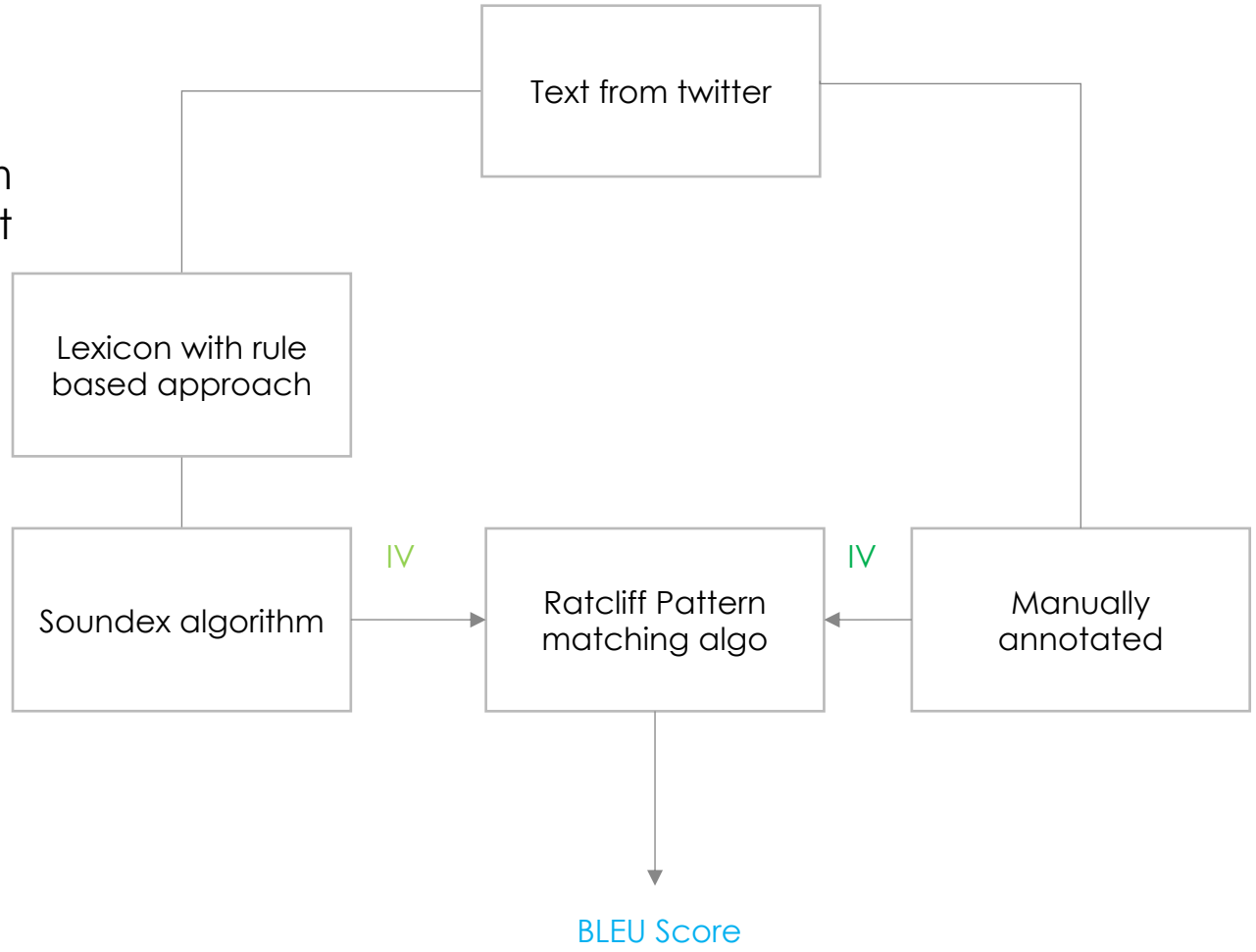
SOUNDEX ENCODING

- Soundex was developed to encode surnames for use in censuses.
- Soundex codes are four-character strings composed of a single letter followed by three numbers.
- Multi word expression is treated as a single atomic unit.

Word	**Soundex Encoding**
abandoned_person	A153_P625
fawn	F500
abandon	A153
abdominal_aortic	A135_A632

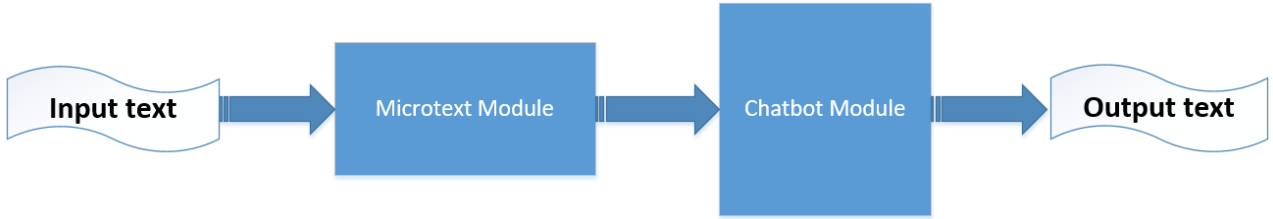
PROPOSED FRAMEWORK

- We used Ratcliff pattern-matching algorithm compares two strings by finding the longest subsequence between them.



APPLICATION TO CHATBOT

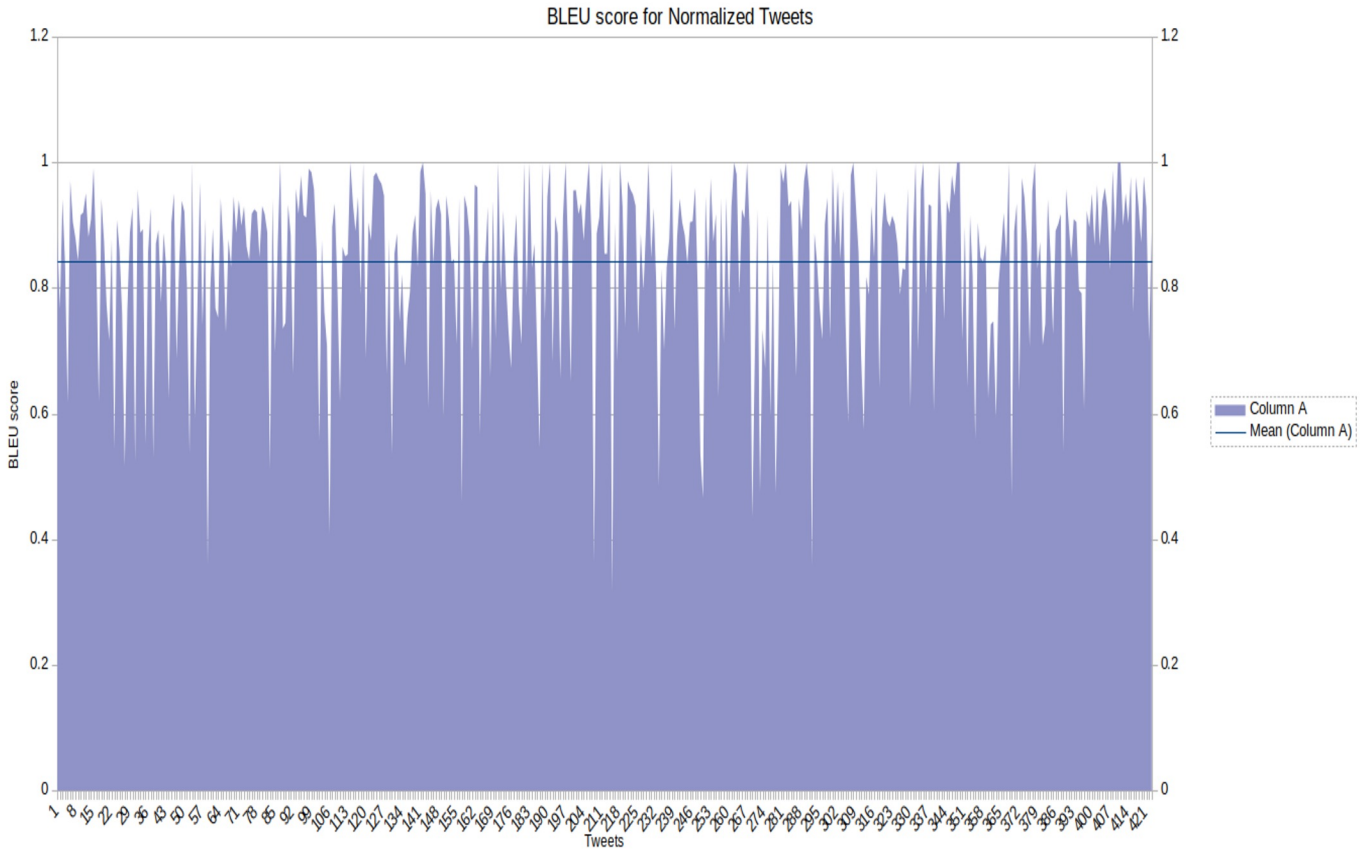
- We used the same normalization model on a chatbot.
- The chatbot is a AI/ML pattern recognition based chatbot [1].
- We trained a binary classifier to detect OOV and IV. Only the OOV were passed through the normalization framework discussed previously.



[1] Ramanathan, Manoj, Nidhi Mishra, and Nadia Magnenat Thalmann. "Nadine Humanoid Social Robotics Platform." In *Computer Graphics International Conference*, pp. 490-496. Springer, Cham, 2019.

RESULT WITH THE TWITTER DATASET ON CHATBOT

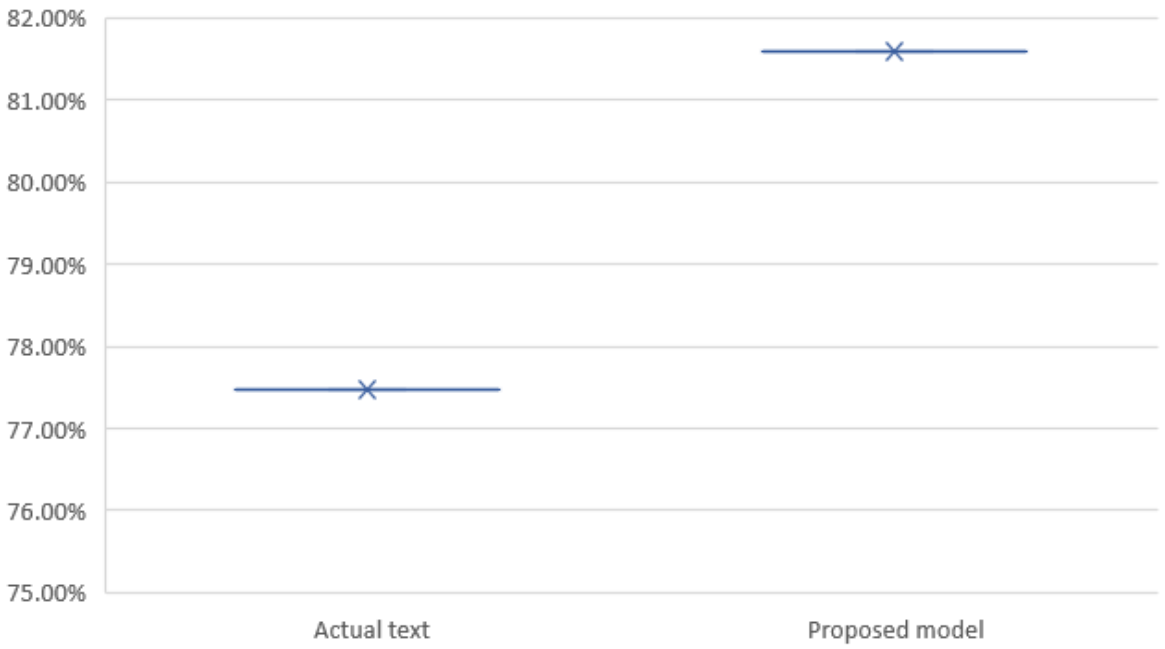
The average BLEU score of tweets show a score of 0.83 after normalization.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

RESULT WITH POLARITY DETECTION

Improvement in polarity detection accuracy by 4% from 77.47% to 81.59%.



SUMMARY

- Soundex algorithm helps in catching the words which might not be present in the lexicon.
- The Ratcliff pattern-matching algorithm compares two strings by finding the longest subsequence between them.
- Improvement in polarity detection accuracy by 4% from 77.47% to 81.59%.
- Improvement in the chatbot understanding for the microtexts.
- 85.31% of texts have a similarity index equal to or greater than 0.8 after normalization.

OUTLINE

- **Introduction**
  - Microtext Processing
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Major Contribution**
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
  - **Major Contribution**
  - **Conclusion and Future work**

IPA BASED MICROTEXT NORMALIZATION

- We used IPA based method and compared our results with existing phonetic based microtext methods on sentiment analysis.
- We built an automatic lexicon method to tackle the microtext normalization method, which can be used on any currently available polarised lexicon, here we have taken SenticNet5 [1].
- The result discusses the impact of coupling sub-symbolic (phonetics) with symbolic (machine learning) Artificial Intelligence to transform the out-of-vocabulary concepts into their standard in-vocabulary form.

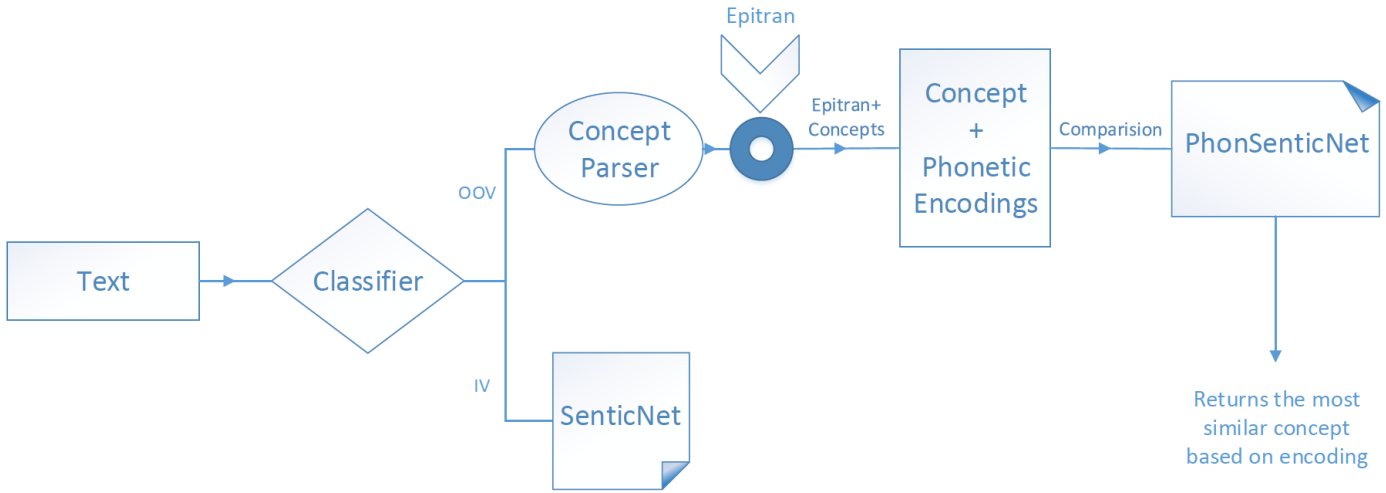
[1] Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018, April). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

IPA BASED MICROTEXT NORMALIZATION

- The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin script. It was devised as a as a standardized representation of speech sounds in written form.
- Epitran [1] is used to transform concepts to their IPA encodings in this work.
- Epitran is a G2P (grapheme-to-phoneme). It takes word tokens in the orthography of a language and outputs a phonemic representation in either IPA.
- It produces precise pronunciations of an utterance which can provide useful feedback on how the model is performing on OOV words during evaluation.

[1] Mortensen, D. R., Dalmia, S., & Littell, P. (2018, May). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

ALGORITHM



Algorithm 1 Algorithm for microtext normalization using phonetic features

Sentence (S) = $s_1, s_2, \cdots, s_n$
c_i = concept-parser(S)
For each concept c_i in Sn
closest-match-concept = PhonSenticNet(c_i)
if closest_ match(C_i, SenticNet) **then**
  return concept polarity
else
  return sentence polarity
end if
average over polarity of concepts for sentence polarity **EndFor**
return sentence polarity

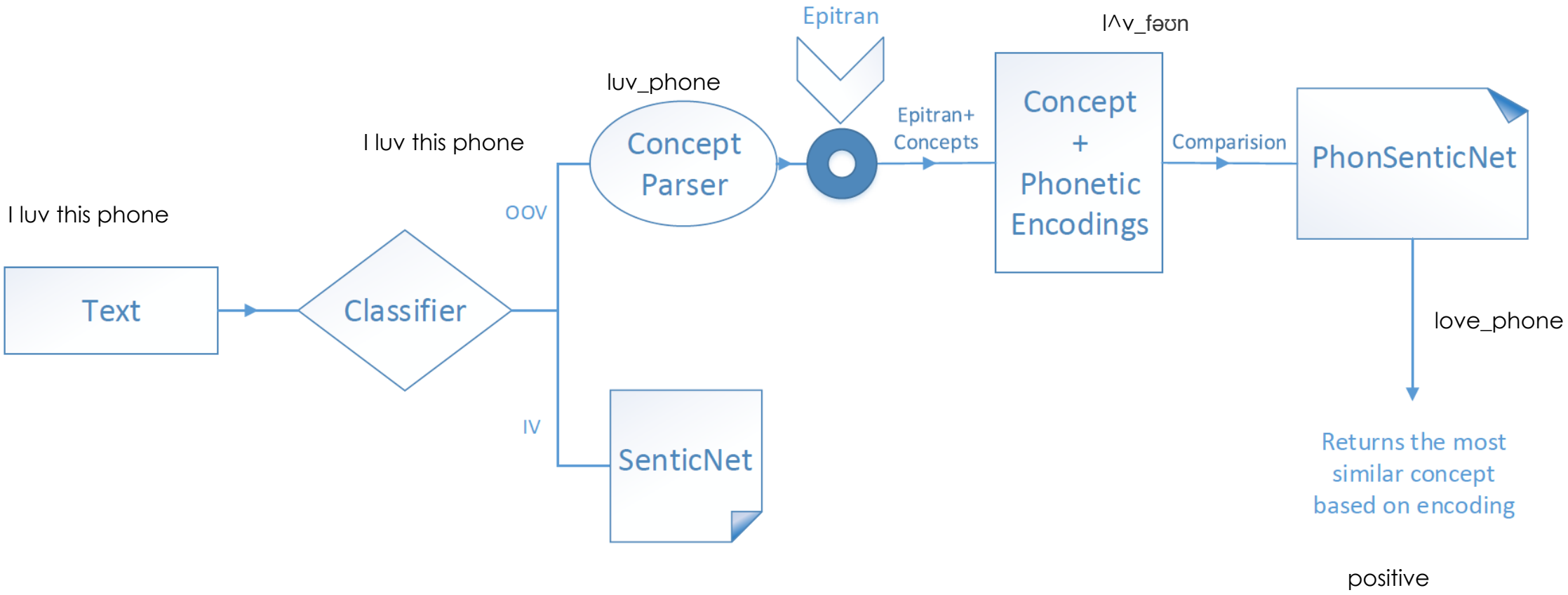
[5]Repetition of a soundex encoding for greater than one

76

Algorithm 2 Closest Match Algorithm

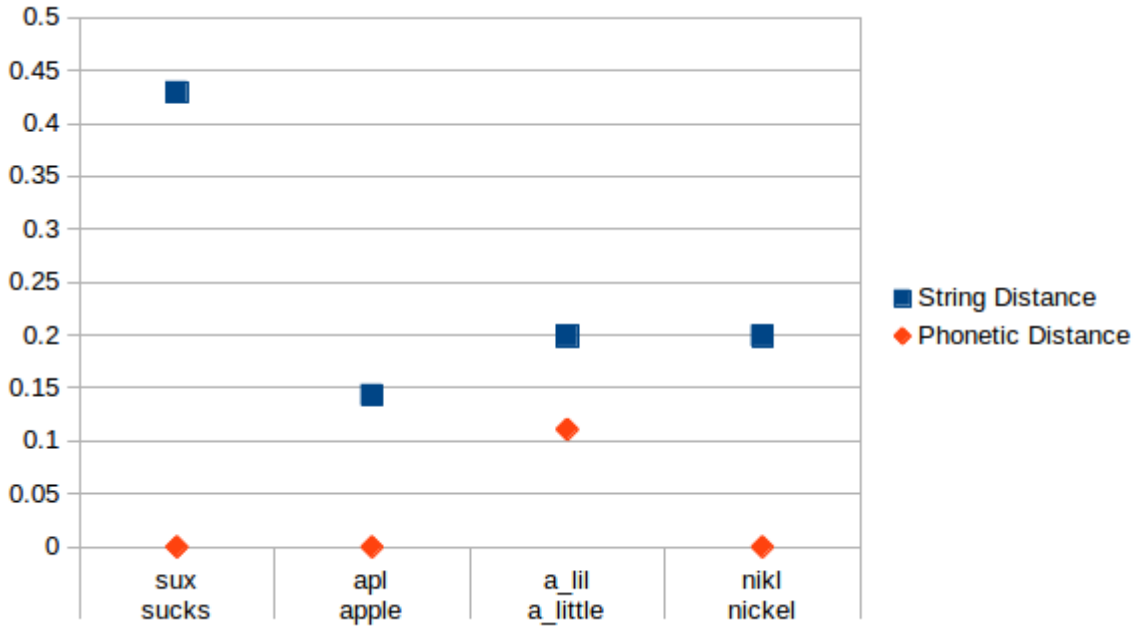
Concept (C) = $c_1, c_2, \cdots, c_m$
For each concept c_i in C
Sorensen (c_i,Senticnet)
EndFor
return phonetically closest matching concept

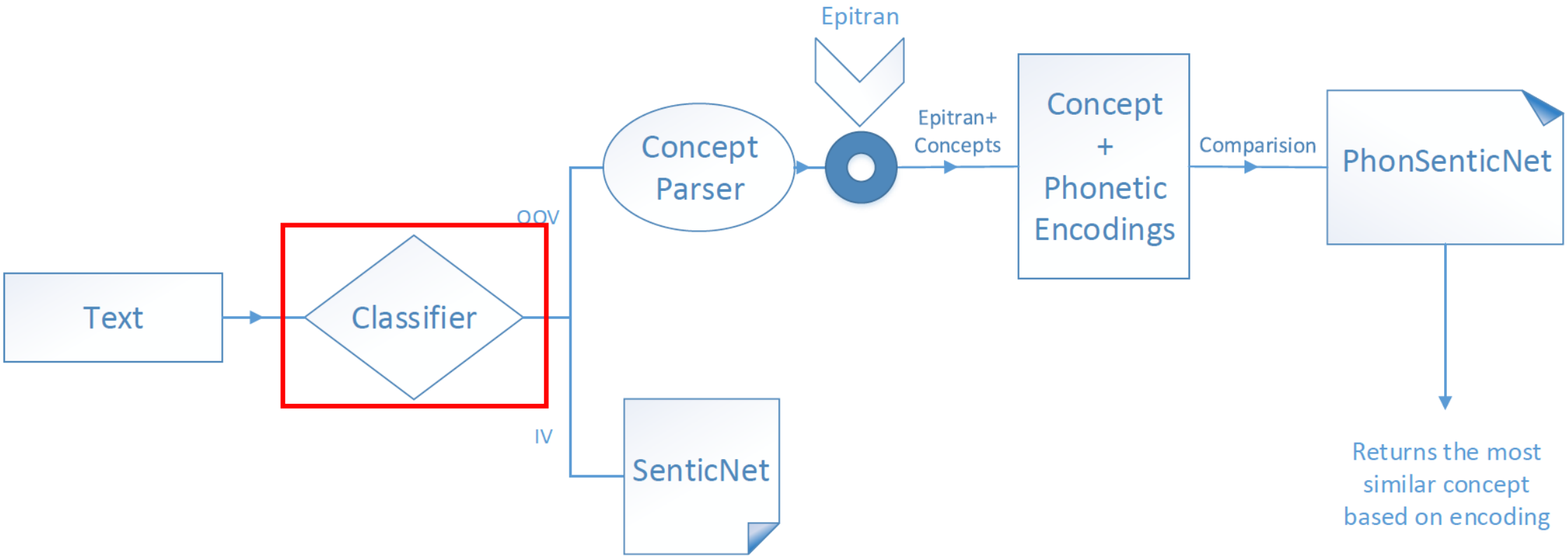
ALGORITHM


I luv this phone
Text
Classifier
I luv this phone
OOV
luv_phone
Concept Parser
IV
SenticNet
Epitran
Epitran+ Concepts
Concept + Phonetic Encodings
lʌv_fəʊn
Comparision
PhonSenticNet
love_phone
Returns the most similar concept based on encoding
positive

EFFECT OF PHONETIC TRANSFORMATION

- Below is the visualisation of string and phonetic distance
- For example:
  - Distance (sux,sucks) = 0.44
  - Distance(IPA(sux), IPA(sucks)) = 0




Epitran
Concept Parser
Epitran+ Concepts
Concept + Phonetic Encodings
Comparision
PhonSenticNet
Text
Classifier
OOV
IV
SenticNet
Returns the most similar concept based on encoding

EVALUATION OF CLASSIFIERS

- Validation on NUS SMS dataset[1]

	Logistic Regression		Stochastic gradient descent		Support vector machine		Multinomial naïve bayes	
	IV	OOV	IV	OOV	IV	OOV	IV	OOV
Precision	91%	95%	84%	98%	87%	97%	89%	97%
Recall	95%	90%	99%	81%	98%	85%	97%	87%
F1	93%	92%	91%	89%	92%	91%	93%	92%
Accuracy	92.75%		89.88%		91.5%		92.25%	

[1] P. Wang and H. T. Ng, "A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation." in HLT-NAACL, 2013, pp. 471–481.

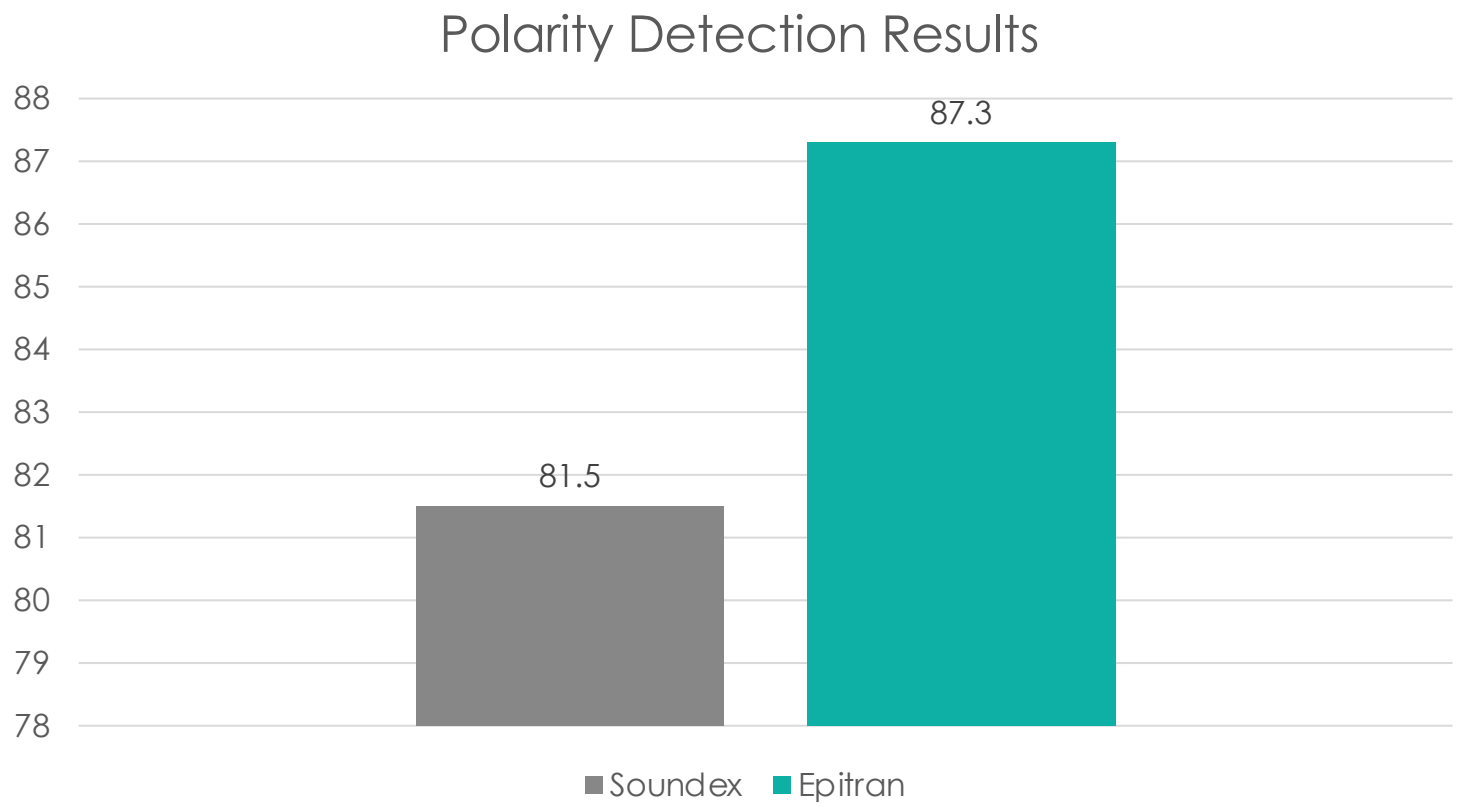
EVALUATION OF CLASSIFIERS

- Validation on Tweets dataset

	Logistic Regression		Stochastic gradient descent		Support vector machine		Multinomial naïve bayes	
	IV	OOV	IV	OOV	IV	OOV	IV	OOV
Precision	71%	69%	63%	72%	74%	67%	81%	68%
Recall	68%	71%	80%	52%	64%	77%	61%	85%
F1	69%	70%	70%	60%	68%	72%	69%	76%%
Accuracy	69.62%		66.05%		70.13%		72.88%	

RESULTS

The accuracy of polarity detection by utilising PhonSenticNet, increases significantly by 6% as shown in Figure below.



SUMMARY

- The proposed framework contains concepts from SenticNet5 and their phonetics which is interpreted using Epitran.
- The lexicon is called as PhonSenticNet which is used as a lexicon for concept-level microtext normalization.
- The input sentence is broken down into concepts and then transformed into their phonetic encoding. The phonetic encoding is matched with the PhonSenticNet.
- Then the most similar matching concept and its corresponding polarity is returned as depicted in the algorithm 1 and 2

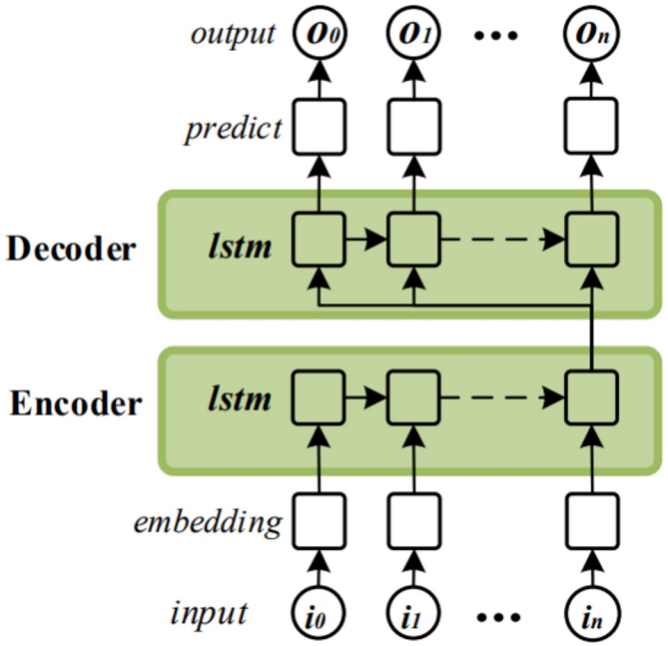
OUTLINE

- **Introduction**
  - Microtext Processing
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Major Contribution**
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
  - Major Contribution
  - **Conclusion and Future work**

SEQ2SEQ MODEL:
FRAMEWORK

Models used:

- LSTM [1,2]
- GRU cell [3]
- CNN with LSTM
- Attentive LSTM



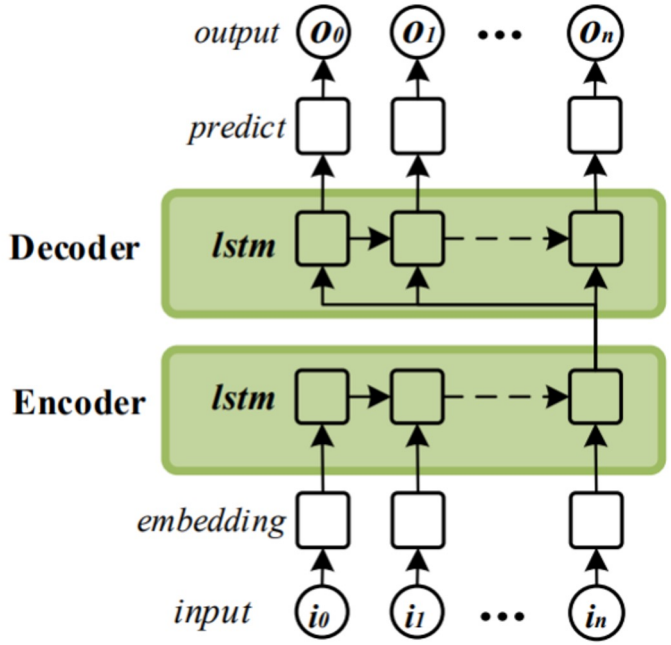
Encoder-Decoder model for Microtext Normalization

[1] Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.
[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735– 1780, 1997
[3]] Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103-111. 2014.

SEQ2SEQ MODEL:
FRAMEWORK

Datasets used:

- Microtext Lexicon
- Tweets
- NUS SMS data [1]
- CMU dictionary [2]
- LexNorm [3]
- CEMt



Encoder-Decoder model for Microtext Normalization

[1] P. Wang and H. T. Ng, "A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation." in HLT-NAACL, 2013, pp. 471–481.
[2] The CMU Pronouncing Dictionary
[3] T. Baldwin, M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, and W. Xu, "Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition," in Proceedings of the Workshop on Noisy User-generated Text, 2015, pp. 126–135

SEQ2SEQ ENCODER DECODER MODEL

- The proposed framework for normalization consists of two parts: an **encoder** for reading the input sequence and encoding it into a fixed length vector, and a **decoder** for decoding the fixed-length vector and outputting the predicted sequence.
- The task of translating OOV words can be understood from the perspective of machine learning as learning the conditional distribution p(w | s) of a target w given a source s.

SEQ2SEQ MODEL:
PARAMETER SETTING

- We used a **2-layer encoder** that reads input characters and a **2-layer decoder** that produces word sequences
- Loss: **L2 loss** for encoder and **ReLu loss** for decoder
- For training: **AdamOptimizer** and learning rate = **.0001**
- Batch size is set to 64 and latent dimension = 256.

SEQ2SEQ MODEL:
CHARACTER TO WORD TRANSLATION MODEL

$$p(\mathbf{w} \mid \mathbf{s}, \theta) = \prod_{j=1}^{m} p(w_j \mid w_{<j}, \mathbf{s}),$$

where θ represents a set of all model parameters in the encoder-decoder model, which are determined by the parameter-estimation process of a standard softmax cross-entropy loss minimization using training data.

Therefore, given θ and s, our normalization task is defined as finding w with maximum probability given by

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{s}, \theta),$$

where $\delta(\cdot)$ denotes the softmax function, w_t refers to the word at time-step t, $\varphi(x_t)$ represents the encoding value from the character at time-step t, and h_{t-1}

$$p(w_t|\cdot) = \delta(\phi(x_t), h_{t-1})$$

SEQ2SEQ MODEL:
CHARACTER TO WORD TRANSLATION MODEL

Cross Entropy Loss

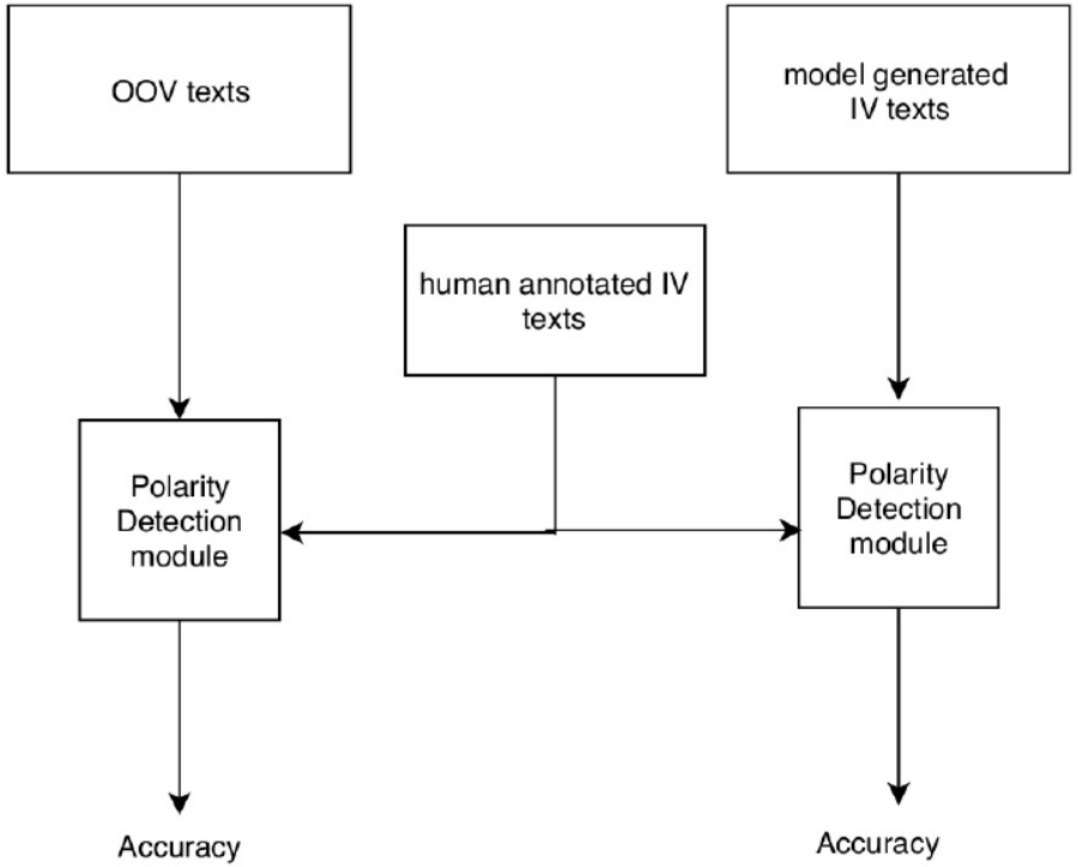
$$L = -\frac{1}{m}\sum_{i}^{m} w_i \log p(w_i, \theta)$$

where m in the denominator is the output length, θ represents a set of all model parameters in the encoder-decoder model.

SEQ2SEQ MODEL:
RESULTS ON MICROTEXT NORMALIZATION

Cases	Datasets	LSTM (%)	Attentive LSTM (%)	GRU cell (%)	CNN with LSTM (%)
With Mask	Lexicon	77.36	**80.63**	78.40	77.82
	Tweets	64.71	64.04	69.85	**69.90**
	SMS	76.15	76.84	**77.24**	77.20
	CMUdict	82.78	82.17	**85.21**	84.73
	LexNorm	74.91	**79.91**	68.71	78.83
	CEMt	64.71	63.84	70.85	**72.86**
Without Mask	Lexicon	78.58	**80.33**	77.22	78.55
	Tweets	65.71	66.93	65.85	**68.55**
	SMS	76.06	**77.09**	76.60	76.59
	CMUdict	81.16	81.10	85.75	**86.87**
	LexNorm	75.8	**80.1**	68.81	79.1
	CEMt	65.55	65.24	68.72	**73.2**

SEQ2SEQ MODEL:
RESULTS ON POLARITY DETECTION



Dataset	OOV Output	Models			
		LSTM	Attentive LSTM	GRU Cell	CNN with LSTM
nus_sms_data	78%	80.4%	82.4%	82.8%	81.3%
norm_tweets	77.47%	79.8%	83.5%	83.8%	81.5%

SEQ2SEQ MODEL:
SUMMARY

- Attentive LSTM works well for short texts.
- The GRU cell model works better for both longer and shorter sentences, as well as for words with mask, yet CNN with LSTM work better without mask.
- **GRU cell** and **CNN with LSTM** both demonstrate the capability to be used in microtext normalization in all the 6 datasets.
- Results show that the attentive LSTM and GRU cell both improve the sentiment analysis accuracy in the range of 4% - 7%. Though LSTM and CNN with LSTM improves the accuracy in the range of 2% - 4%.

OUTLINE

- **Introduction**
  - Microtext Processing
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
  - **Major Contribution**
  - Conclusion and Future work

MAJOR CONTRIBUTIONS

- We built a framework with **microtext lexicon, a rule-based method and Soundex algorithm** for microtext.
- We found that there is a definite impact of normalization to the **existing sentiment analysis framework and a chatbot**.
- The proposed method was documented and published in:
  - Satapathy Ranjan, Claudia Guerreiro, Iti Chaturvedi, and Erik Cambria. "Phonetic based microtext normalization for twitter sentiment analysis." In ICDMW, pp. 407-413. IEEE, 2017.

MAJOR CONTRIBUTIONS

- We built a framework for **phonetic based microtext normalization** by transforming the tokens to phonetic subspace.
- We used **IPA based method** and compared our results with existing phonetic based microtext methods. We built an automatic lexicon method to tackle the microtext normalization method, which can be used **on any currently available polarised lexicon**.
- The proposed method was documented and published in:
  - Satapathy, Ranjan, Aalind Singh, and Erik Cambria. "PhonSenticNet: A cognitive approach to microtext normalization for concept-level sentiment analysis." In International Conference on Computational Data and Social Networks, pp. 177-188. Springer, Cham, 2019.

MAJOR CONTRIBUTIONS

- We built a seq2seq based Deep Learning method for transforming **out-of-vocabulary** tokens to **in-vocabulary** tokens.
- We developed baselines for Deep Learning based methods for microtext normalization.
- The proposed method was documented and published in:
  - Satapathy, Ranjan, Yang Li, Sandro Cavallari, and Erik Cambria. "Seq2seq deep learning models for microtext normalization." In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2019.

OUTLINE

- **Introduction**
  - Microtext Processing
  - Features of Microtext
  - Motivation
- **Literature Survey**
  - Related work to Microtext Normalization
- **Current Work**
  - Unsupervised learning based method
    - Hybrid method
    - IPA based method
  - Supervised learning based method
    - Seq2Seq method
  - Major Contribution
  - **Conclusion and Future work**

CONCLUSION

- Language imposes constraints on the meaning of a sentence, both on the basis of the syntactic form of the sentence and on the basis of its contextual usage.
- Knowledge of these linguistic constraints can help us design NLP systems with higher accuracies.
- Texting language involves the phonetics and abbreviations, which requires linguistic knowledge to be incorporated into the model.
- Therefore, this thesis studies the paradigm shift from the standard language to texting language .

CONCLUSION

- We propose an **unsupervised hybrid method** to transform OOV sentences to IV sentences. The transformed sentences are then passed through a polarity detection module and compared with human-annotated IV sentences for polarity.
- To tackle the phonetic replacement challenge posed, we develop a framework for **IPA based** OOV concepts to IV concepts. We propose concept-level microtext normalization algorithm in this regard, to transform the OOV concept or word to its phonetic subspace. However, as the language is very dynamic, it is cumbersome to maintain a lexicon.
- We constructed a **Deep Learning based framework**, to transform OOV to IV. We showed different variations of Deep Learning models on 6 different datasets.

FUTURE WORK

- Driverless Cars [1]

- Code switching and code mixing [2]

[1] *Lipson, Hod, and Melba Kurman. Driverless: intelligent cars and the road ahead. Mit Press, 2016..*
[2] Myers-Scotton, Carol. "Code-switching." *The handbook of sociolinguistics* (2017): 217-237.

SUMMARY

- Introduction to Microtext
- Classes of Microtext
- Problems arising due to Microtexts
- Major contribution
- Conclusion and Future work

PUBLICATIONS

CONFERENCE

- **Ranjan Satapathy**, Claudia Guerreiro, Iti Chaturvedi, and Erik Cambria. "Phonetic-based microtext normalization for twitter sentiment analysis." In 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 407-413. IEEE, 2017. •
- **Ranjan Satapathy**, Yang Li, Sandro Cavallari, and Erik Cambria. "Seq2Seq Deep Learning Models for Microtext Normalization" In International Joint Conference on Neural Networks (IJCNN), pp.1–8, 2019.
- **Satapathy R.**, Singh A., Cambria E. (2019) PhonSenticNet: A Cognitive Approach to Microtext Normalization for Concept-Level Sentiment Analysis. In: Tagarelli A., Tong H. (eds) Computational Data and Social Networks. CSoNet 2019. Lecture Notes in Computer Science, vol 11917. Springer, Cham.
- David Vilares, Haiyun Peng, **Ranjan Satapathy**, and Erik Cambria. "BabelSenticNet: a commonsense reasoning framework for multilingual sentiment analysis." In 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1292-1298. IEEE, 2018.
- Mishra, N., Ramanathan, M., **Satapathy, R.**, Cambria, E., & Magnenat-Thalmann, N. Can a Humanoid Robot be part of the Organizational Workforce? A User Study Leveraging Sentiment Analysis. In 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN, 2019) (pp. 1-7). IEEE.

JOURNAL

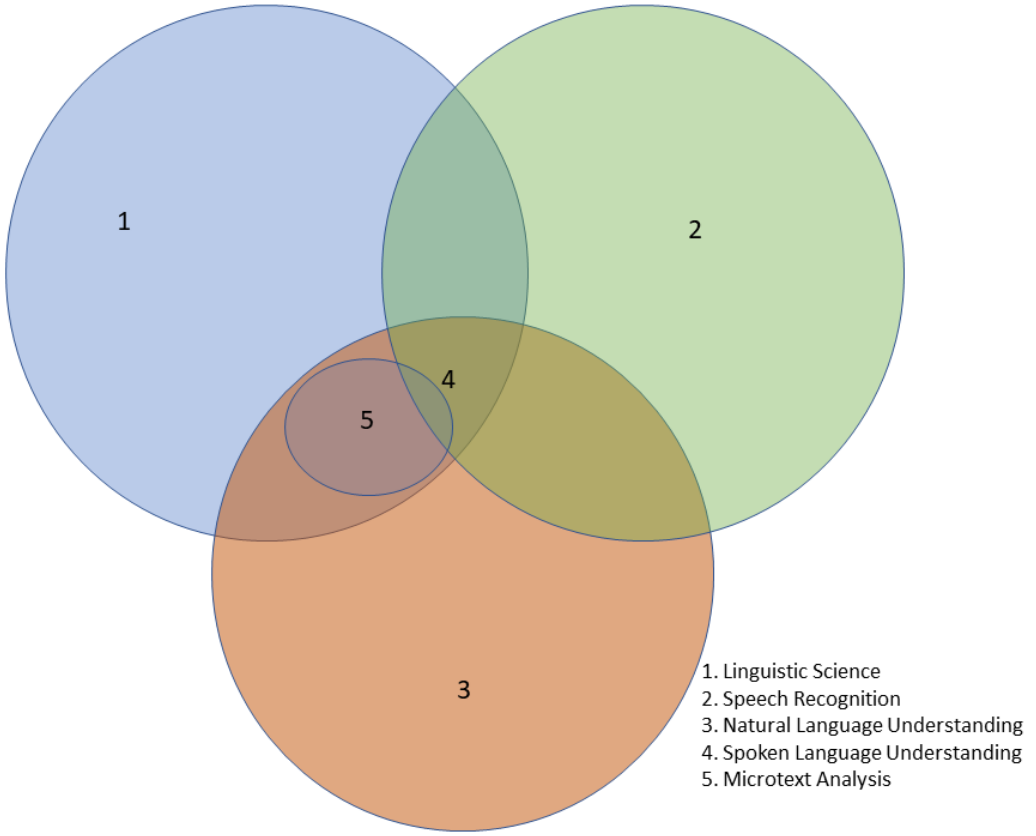
- **Satapathy, R.**, Cambria, E., Nanetti, A. , Hussain, A,. A Review of Shorthand Systems: From Brachygraphy to Microtext and Beyond. Cogn Comput 12, 778–792 (2020).
- **Ranjan Satapathy**, Iti Chaturvedi, Erik Cambria, Shirley S. Ho, and Jin Cheon Na. "Subjectivity detection in nuclear energy tweets." Computaci´on y Sistemas 21, no. 4 (2017): 657-664.
- Iti Chaturvedi, **Ranjan Satapathy**, Sandro Cavallari, and Erik Cambria. "Fuzzy commonsense reasoning for multimodal sentiment analysis." Pattern Recognition Letters 125 (2019): 264-270.

Thank You

ADDITIONAL SLIDES

INTRODUCTION:
MICROTEXT PROCESSING

- The term "Microtext Analysis" refers to the branch of **Natural Language Processing** (NLP) that focuses on handling "semi-structured" texts.
- It utilizes knowledge from different domains.
- Users exhibit different amounts of shortened English terms and different shortening styles depending upon the geolocation and culture[1] they are influenced by.



[1] Gouws S, Metzler D, Cai C, Hovy E. Contextual bearing on linguistic variation in social media. In: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics; 2011. p. 20–29

CORPUS FOR ENGLISH MICROTEXT NORMALIZATION (CEMT-NORM)

- We describe a methodology to build the corpus with guidelines followed during the annotation process. We also present some baseline results of the corpus on machine learning algorithms and make the code and resources publicly available.
- A total of 1432 OOV sentences and their IV sentences made it to the final corpus with their corresponding polarity. We also show the accuracy of character-level sentiment analysis module as a baseline.

PREPARATION OF CORPUS

- We crawled twitter to get tweets, with no constraint on keywords.
- We passed the tweets through pre-processing pipeline.
- Next, we employ three annotators to clean each tweet and write it's corresponding in-vocabulary(IV) text.
- It followed a majority voting to decide the final tweet and its IV text.
- We dropped the IV text where any of the three annotations were different, and also remove their corresponding tweet from the corpus to maintain coherence

PRE-PROCESSING STEP

- We used the basic modules such as a dictionary, affix analysis, number and dates detection to analyse tokens in the corpus, and a token is considered as an OOV if there is no match in any of the modules.
- As a first step, we tokenized the sentences to get tokens.
- The tokenizer's rules were tuned to remove usernames (), hashtags (#), e-mail addresses, URLs, and emoticons.
- As a second pre-processing step, we used Pyenchant to identify OOV words in the corpus.
- The dataset contains sentences with at least one OOV word.

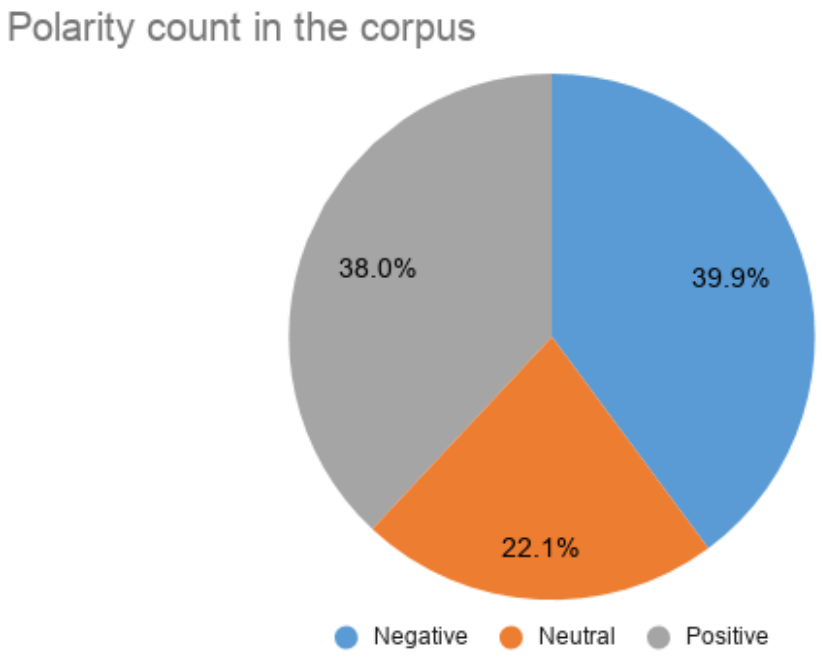
CORPUS STATISTICS

Average unique words per sentence : 4679/1432 ≈ 3.3
Average unique OOV words per sentence : 1861/1432 ≈ 1.3

	Total vocabulary	Unique
OOV	3541	1861
Words	15798	4679

- The corpus contains around 1432 text and is annotated by 3 experts (Cohen's kappa = 0.79) in terms of polarity (positive, negative or neutral).
- We also asked 3 independent experts to normalize each of the OOV text to IV text

CORPUS STATISTICS



- The corpus has 1432 OOV text with their corresponding IV text. The corpus also contains sentiment of the text.
- We calculated the OOV words using Pyenchant tool as a dictionary. The words which return False were regarded as OOV.
- We also calculated the unique set of vocabulary for both total words and OOV.
- The unique words in total vocabulary are 4679, and unique OOV words are 1861, which is about 40% of total unique words

REFERENCES

[1] T.~Matsui, S.~Furui, Concatenated phoneme models for text-variable speaker recognition, in: Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, Vol.~2, IEEE, 1993, pp. 391--394.

[2] S.~Bartlett, G.~Kondrak, C.~Cherry, Automatic syllabification with structured svms for letter-to-phoneme conversion, Proceedings of ACL-08: HLT (2008) 568--576.

[3] C.~Zhang, T.~Baldwin, H.~Ho, B.~Kimelfeld, Y.~Li, Adaptive parser-centric text normalization, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol.~1, 2013, pp. 1159--1168.

[4] S.~Brody, N.~Diakopoulos, Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs, in: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, 2011, pp. 562--570.

[5] P.~Wang, H.~T. Ng, A beam-search decoder for normalization of social media text with application to machine translation., in: HLT-NAACL, 2013, pp. 471--481.

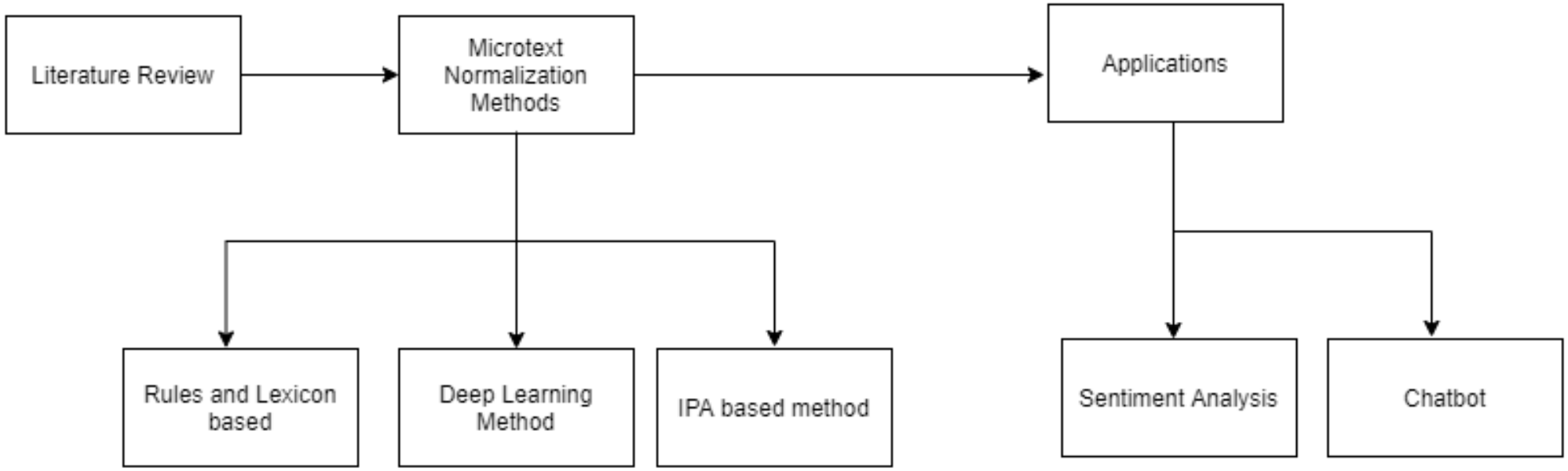
[6] M.~Zare, S.~Rohatgi, Deepnorm-a deep learning approach to text normalization, arXiv preprint arXiv:1712.06994.

REFERENCES

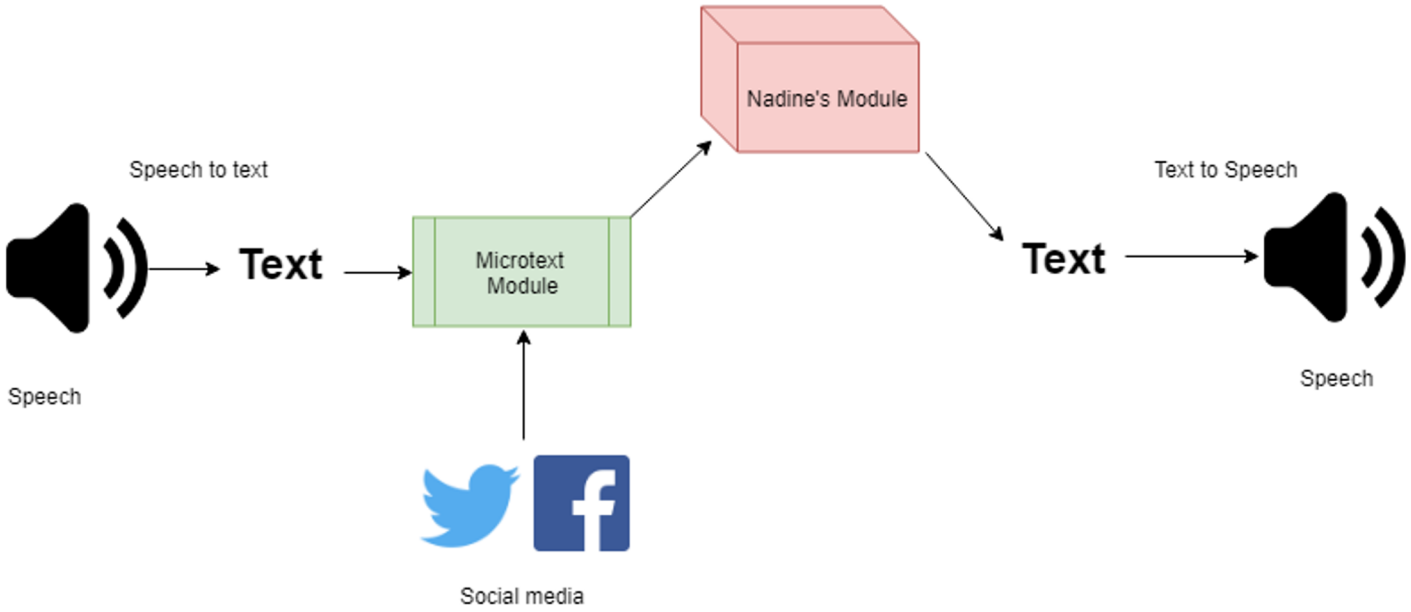
[7] Reut Tsarfaty, Djame Seddah, Sandra K ´ ubler, and ¨ Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. Computational Linguistics, 39(1):15–22.

[8] Ruhi Sarikaya, Katrin Kirchhoff, Tanja Schultz, and Dilek Hakkani-Tur. 2009. Introduction to the special issue on processing morphologically rich languages. Trans. Audio, Speech and Lang. Proc., 17(5):861–862, July.

CONCLUSION

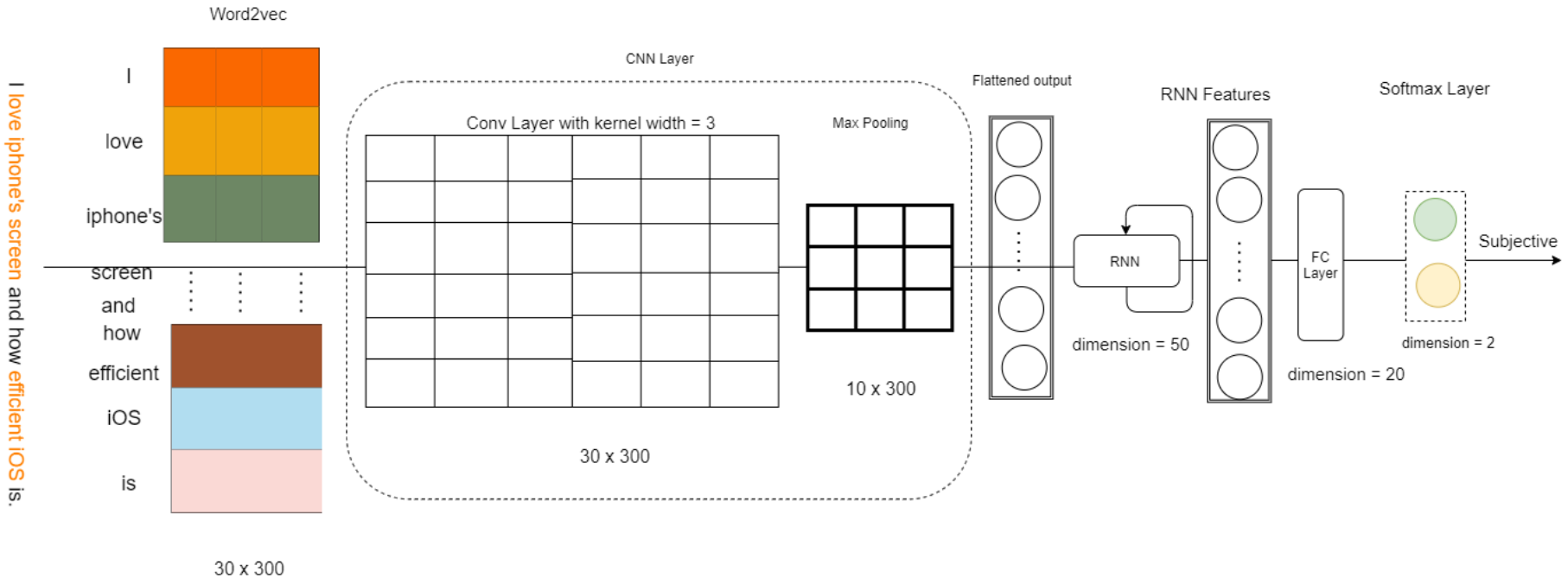

Literature Review
Microtext Normalization Methods
Applications
Rules and Lexicon based
Deep Learning Method
IPA based method
Sentiment Analysis
Chatbot

CHATBOT FRAMEWORK



Framework for Nadine's System

MOTIVATION (THIS MODEL SHOULD BE MOVED TO THE END)



MOTIVATION (THIS MODEL SHOULD BE MOVED TO THE END)

	Framework	MPQA dataset (F1-measure)	Nuclear Energy Tweets (F1-measure)	Silk Road Tweets (F1-measure)
Baselines	CNN	89.4%	55.6%	56.3%
	GBN + CNN	93.2%	57.4%	57.1%
	GBN + CNN + RNN	97.4%	58.9%	59.4%
Proposed Model	GBN + CNN + RNN+ RL	-	59.7%	61.5%

IMPORTANCE OF PHONETIC FEATURES

```
  if np.issubdtype(vec.dtype, np.int):
Word similar to luv: partysquad, Similarity: 0.58
Word similar to luv: girlz, Similarity: 0.57
Word similar to luv: funkalicious, Similarity: 0.55
Word similar to luv: supastar, Similarity: 0.55
Word similar to luv: supastarr, Similarity: 0.55
Word similar to luv: brillz, Similarity: 0.55
Word similar to luv: smoove, Similarity: 0.55
Word similar to luv: twerk, Similarity: 0.54
Word similar to luv: love×trax, Similarity: 0.54
Word similar to luv: remixx, Similarity: 0.54
Word similar to thr: thrr, Similarity: 0.59
Word similar to thr: thrrr, Similarity: 0.59
Word similar to thr: othrr, Similarity: 0.57
Word similar to thr: srrn, Similarity: 0.56
Word similar to thr: imagr, Similarity: 0.55
Word similar to thr: pagrs, Similarity: 0.54
Word similar to thr: problrm, Similarity: 0.54
Word similar to thr: pagr, Similarity: 0.54
Word similar to thr: articlrs, Similarity: 0.54
Word similar to thr: srrms, Similarity: 0.53
Word similar to wassup: •wassup, Similarity: 0.76
Word similar to wassup: whassup, Similarity: 0.76
Word similar to wassup: nigga, Similarity: 0.59
Word similar to wassup: wassupwestcoast, Similarity: 0.59
Word similar to wassup: wussup, Similarity: 0.58
Word similar to wassup: hey, Similarity: 0.58
Word similar to wassup: jeeez, Similarity: 0.58
Word similar to wassup: wyassup, Similarity: 0.57
Word similar to wassup: yeah, Similarity: 0.57
Word similar to wassup: rappin, Similarity: 0.57
Word similar to girls: boys, Similarity: 0.86
Word similar to girls: schoolgirls, Similarity: 0.74
Word similar to girls: girls,, Similarity: 0.72
Word similar to girls: girls…, Similarity: 0.72
Word similar to girls: boy/girls, Similarity: 0.71
Word similar to girls: girls–and, Similarity: 0.69
Word similar to girls: girls–a, Similarity: 0.68
Word similar to girls: girls`, Similarity: 0.68
Word similar to girls: girl, Similarity: 0.67
Word similar to girls: –girls, Similarity: 0.66
```

Top 10 similar words in Wikipedia dump for some of the microtexts using **FASTTEXT** [luv, thr, wassup, girls]

INTRODUCTION:
MICROTEXT PROCESSING

- Microtext has become one of the most widespread communication forms among users due to its casual writing style and colloquial tone.

- Until May 2020 there were nearly 500 million tweets sent each day, i.e. around 6,000 tweets every second.

O.K. IS A MICROTEXT AND IS CONSISTENTLY USED