SMAI Project

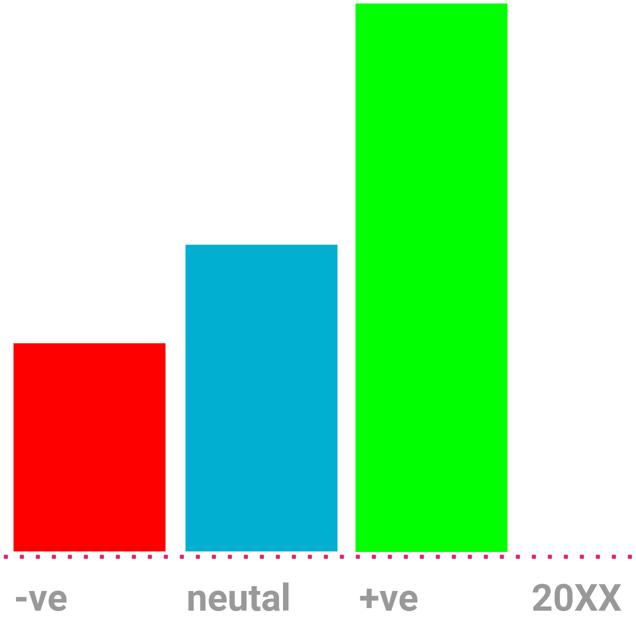
Itisha Rajat Dewan
Shikha Jain
Rashi Shrishrimal

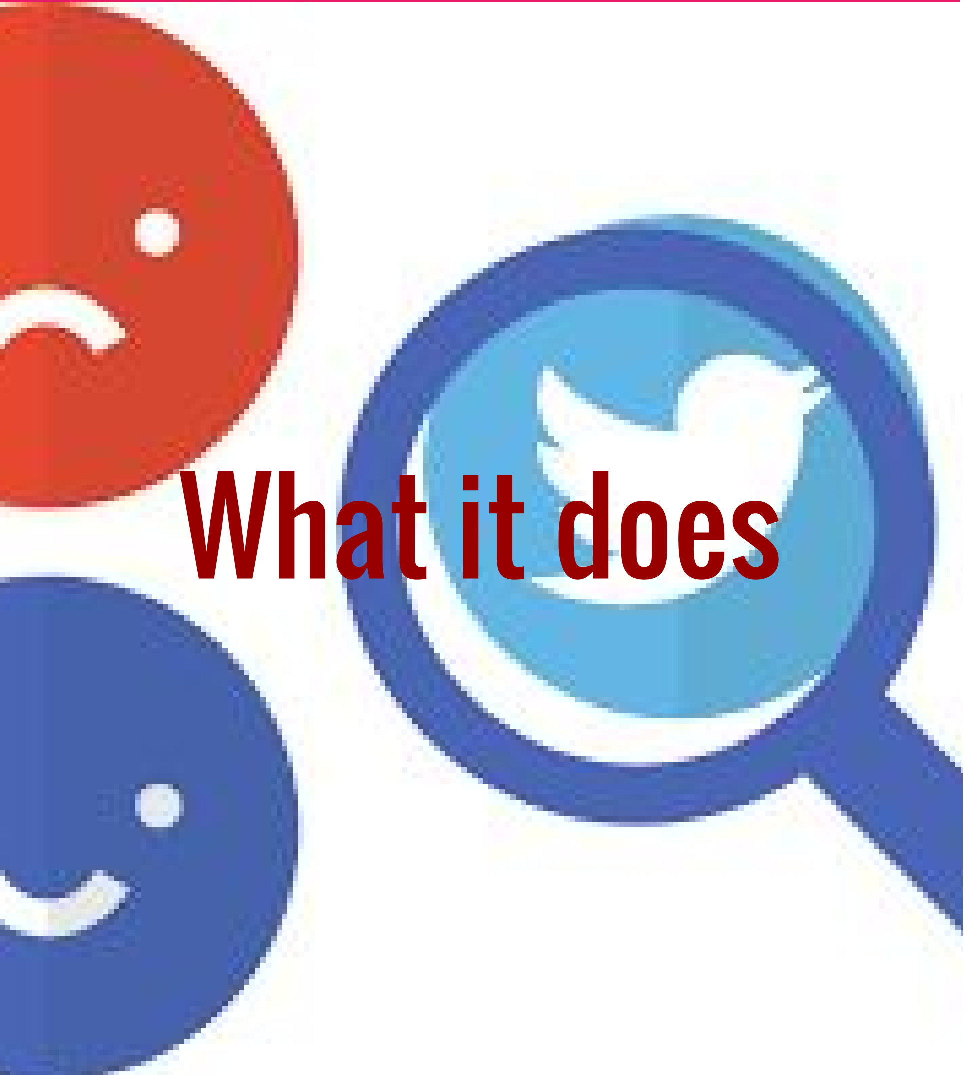
Twitter Sentiment Analysis

input
analysis
correspondence
SentiCorr
expressing
Outlook
patterns
automated
set
multilingual
subjective
problem
e-mail
short
exploration
negative
output
language
tagging
dimensions
social
MS
different
texts
neutral
algorithm
generated
person
POS-Tags
relevance
plug-in
Interface
presented
zoom
objective
languages
sentiment
four-step
emotion
bipolar
developed
positive
learning
interested
software
sources
model
aware
POS
OLAP
separate
news
steps
message
detection
media
LIGA
information
process
Hyves
approach
sentence
feedback
content
text
identification
summary
e-mails
English
polarity
messages
Netherlands
knowledge
provide
Polarity
stress
data
speech
Facebook
AdaBoost
RBEM
current
people
Twitter
subjectivity
written
user
framework
classification

What is Sentiment Analysis ?

The use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.





What it does

It digs deeper into the opinions of millions of users to find out what do they say about you or certain things you want to know!

Data Set

Twitter is a social networking that allows users to post real time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service, people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets.

Target Hashtags Emoticons

The dataset consists of 8222 manually annotated tweets and was tested by using cross validation techniques.

Twitter Data

Answer the question, "Why are we the ones to solve the problem we identified?"

Casey Baumer

Gas by my house hit $3.39!!!! I'm going to Chapel Hill on Sat. :)

Rahul Gupta

@oluoch @victor_otti @kunjand I just watched it! Sridevi's comeback.... U remember her from the 90s?? Sun mornings on NTA ;)

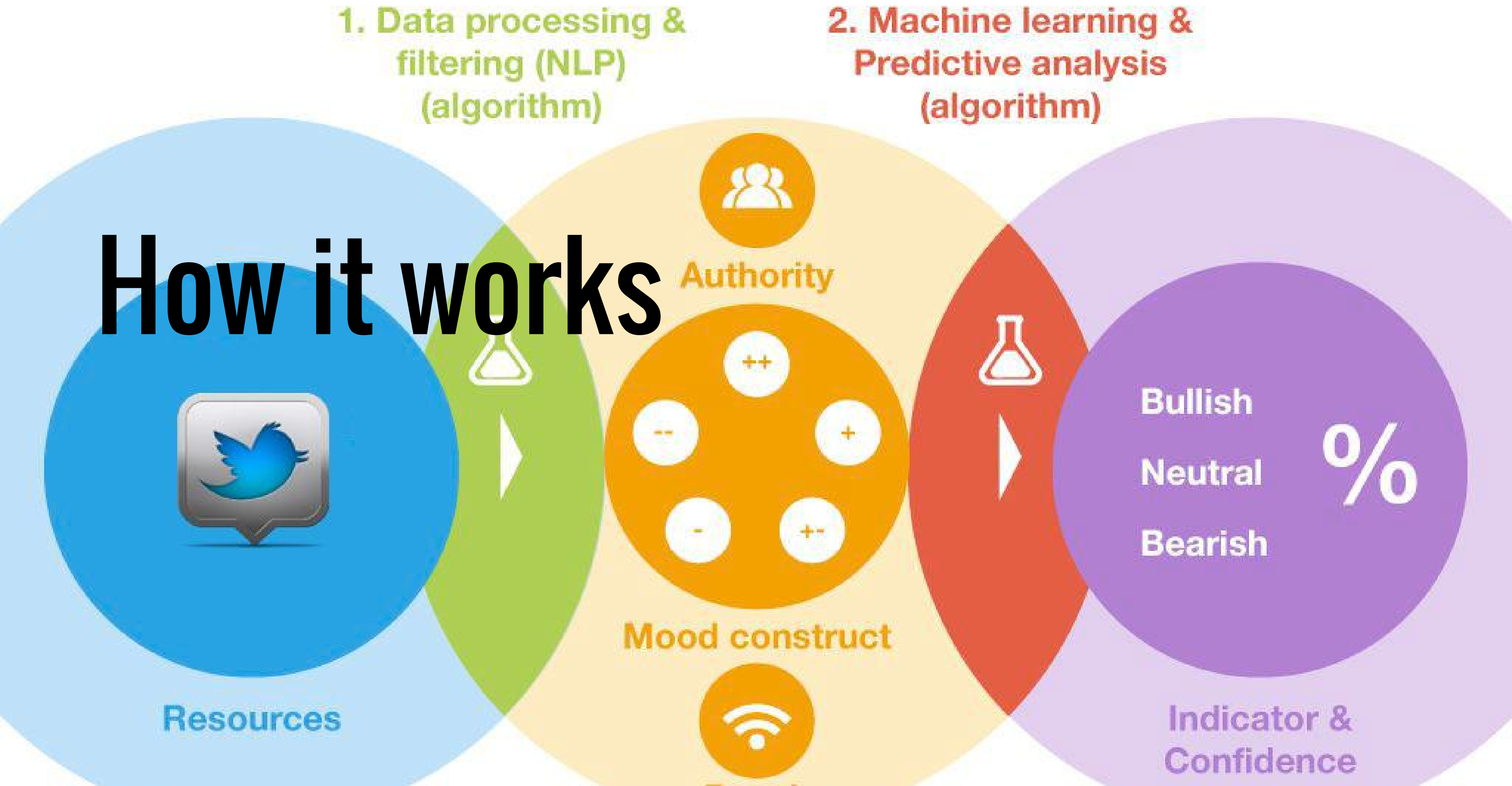
Ashley Wilson

@MsSheLahY I didnt want to just pop up... but yep we have chapel hill next wednesday you should come.. and shes great ill tell her you asked

Jake Tanner

Obma fails 2 unite us, he divides us by sex, color, party, age, wealth Nevada's 1st Latino governor votes early for Romney, predicts GOP win

How it works

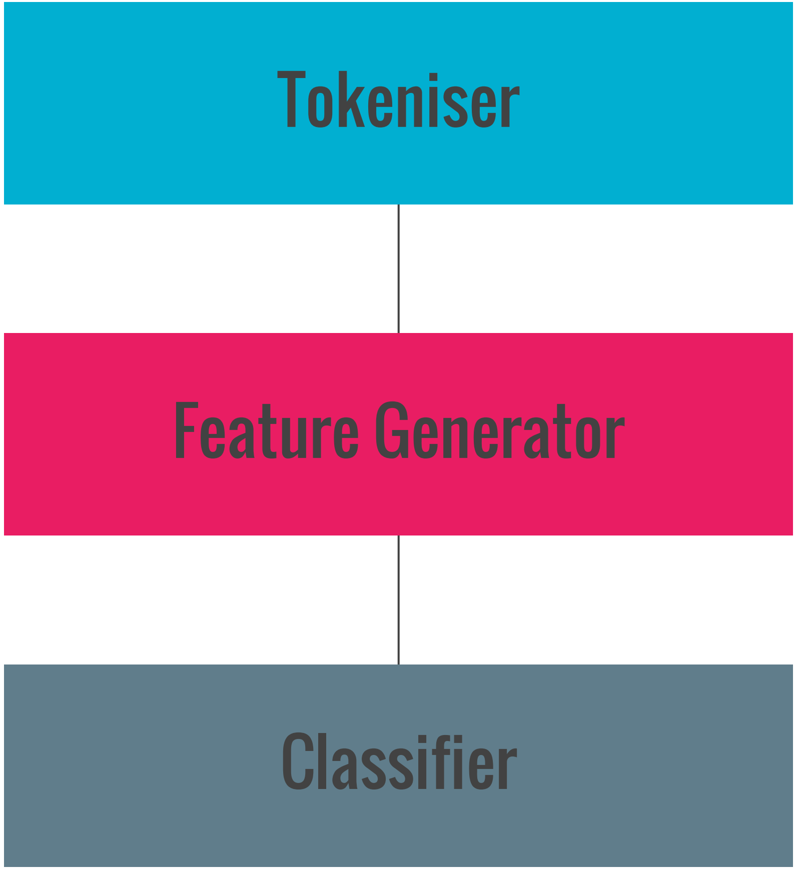

1. Data processing & filtering (NLP) (algorithm)
2. Machine learning & Predictive analysis (algorithm)
Authority
++
--
+
-
+-
Mood construct
Bullish
Neutral
Bearish
%
Resources
Indicator & Confidence

Classification model

Used typical SVM with linear kernel as the classfier.

Generated features were sent to the classifier to model the data.

Used 9 - fold cross validation to validate.



Tokeniser

Feature Generator

Classifier

Preprocessing and Tokenizer

- Lemmatization (removed due to poor accuracy)
- Stop word removal
- Replace all acronyms with their full-forms by looking up at the acronyms dictionary.
- Replace all the emoticons with their sentiment polarity by looking up the emoticon dictionary
- Replace all negations (e.g. not, no, never, n’t, cannot) by tag “NOT”
- Replace all URLs with a tag ||U||
- Replace targets with tag ||T||

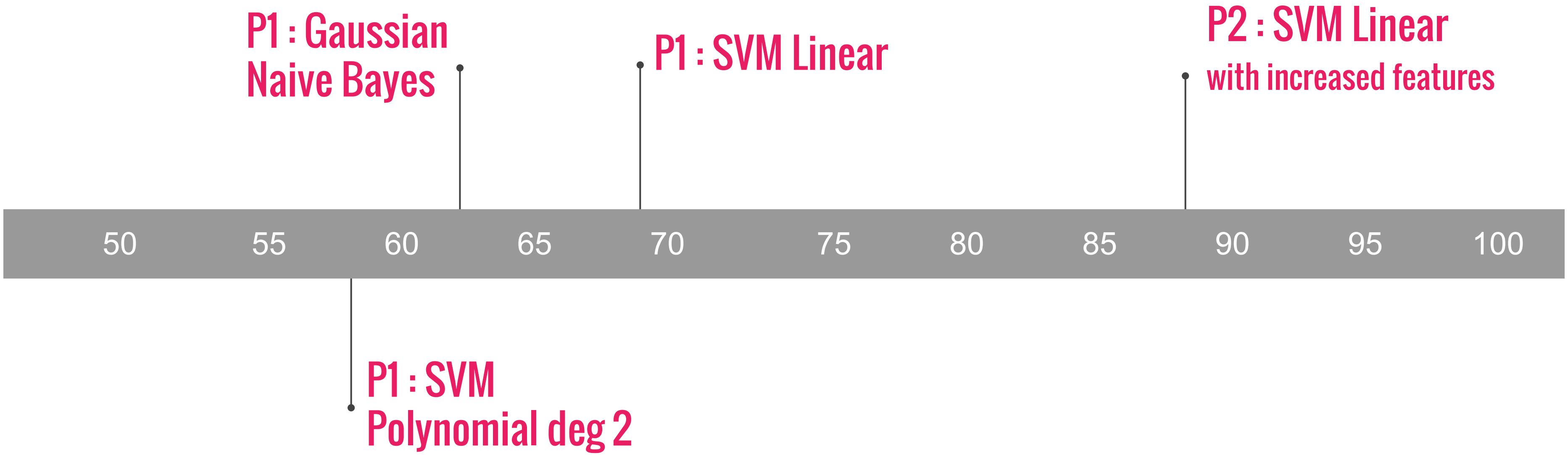
Features Used

- emoticons
- whether the last token is a positive or negative emoticon;
- elongated words
- ends with one of the punctuation marks: ? , !, ‘, “, etc.
- A negated context affects the n-gram and lexicon features:

- word n-grams
- non-contiguous ngrams
- character n-grams
- all-caps
- POS tagging
- hashtags
- Punctuation;

Results

Various Classifiers and accuracies



P1 : Gaussian Naive Bayes

P1 : SVM Linear

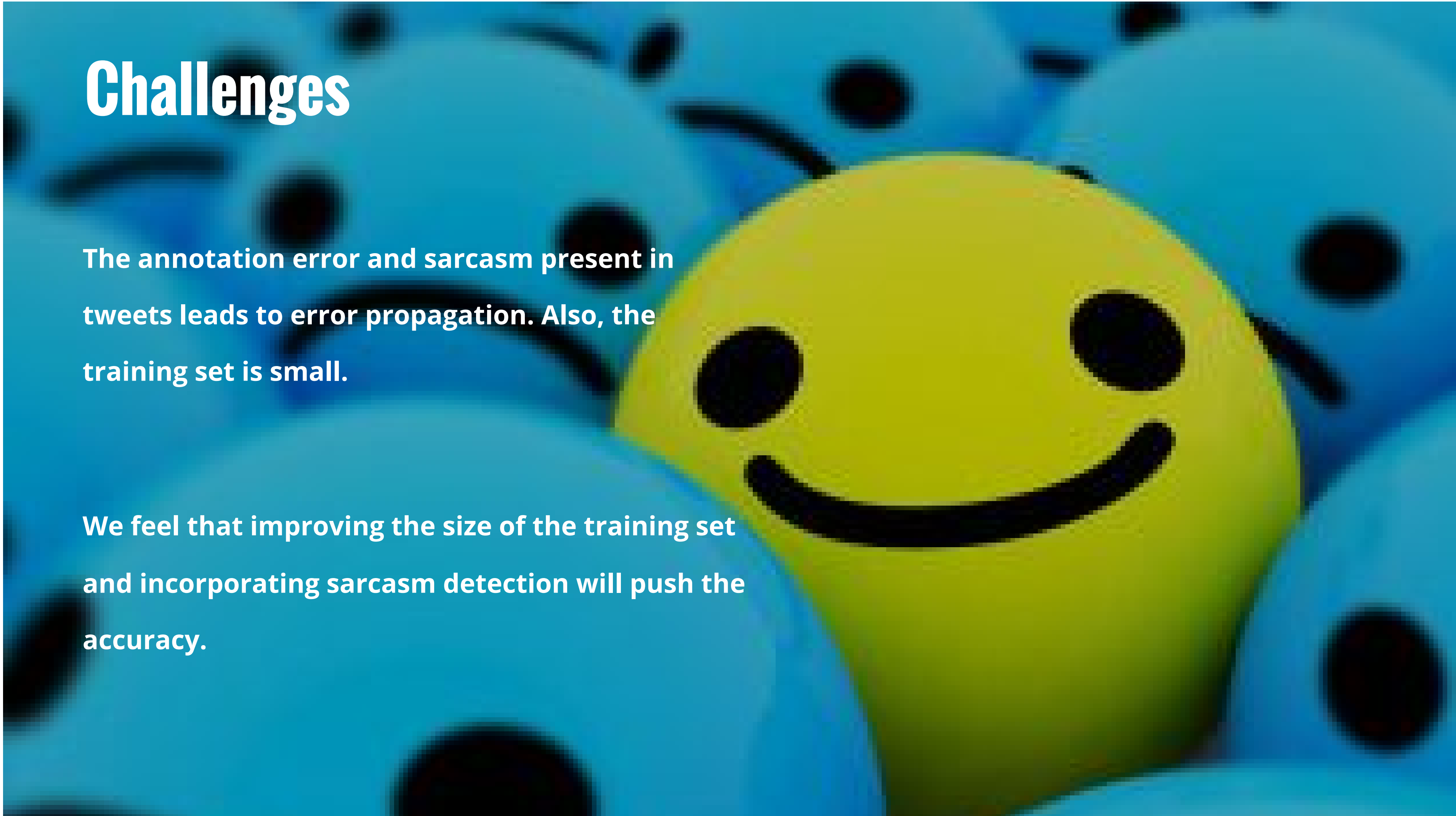
P2 : SVM Linear with increased features

P1 : SVM Polynomial deg 2

Conclusion

We implemented a variety of features based on surface form and lexical categories. The sentiment lexicon features (both manually created and externally generated) along with ngram features (both word and character ngrams) led to the most gain in performance.

Challenges



The annotation error and sarcasm present in tweets leads to error propagation. Also, the training set is small.

We feel that improving the size of the training set and incorporating sarcasm detection will push the accuracy.

Thank You