Semantic Textual Similarity

Unai Gurbindo
Jaume Guasch

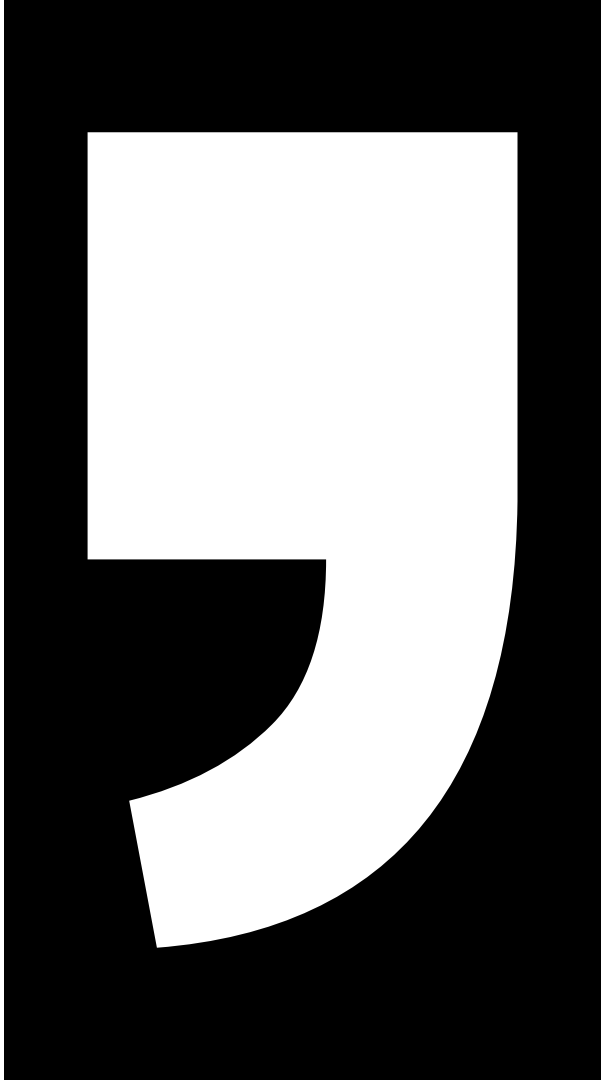


Table of Contents

Introduction

Semantic Textual Similarity task of SemEval-2012 Task 6

$s1 =$ The chef prepared a delicious pasta dish.

$s2 =$ The cook made a tasty pasta meal.

$sim(s1, s2)$?

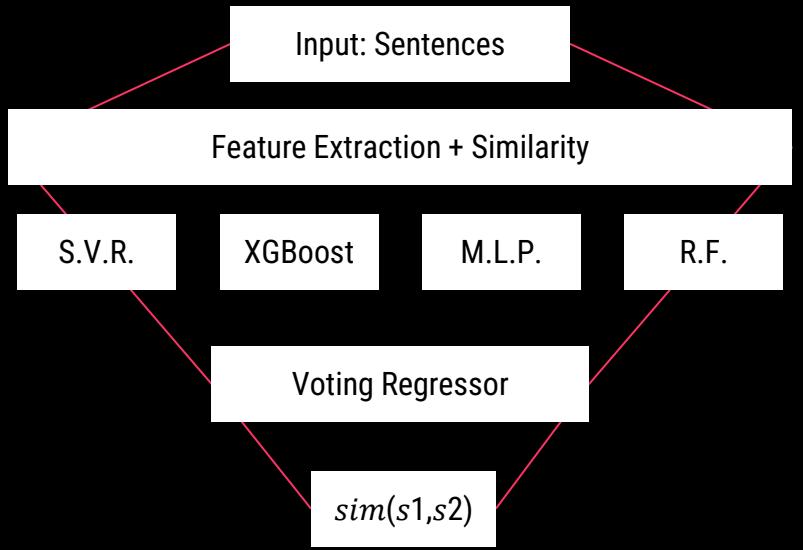


Table of Contents

Table of Contents

Preprocessing

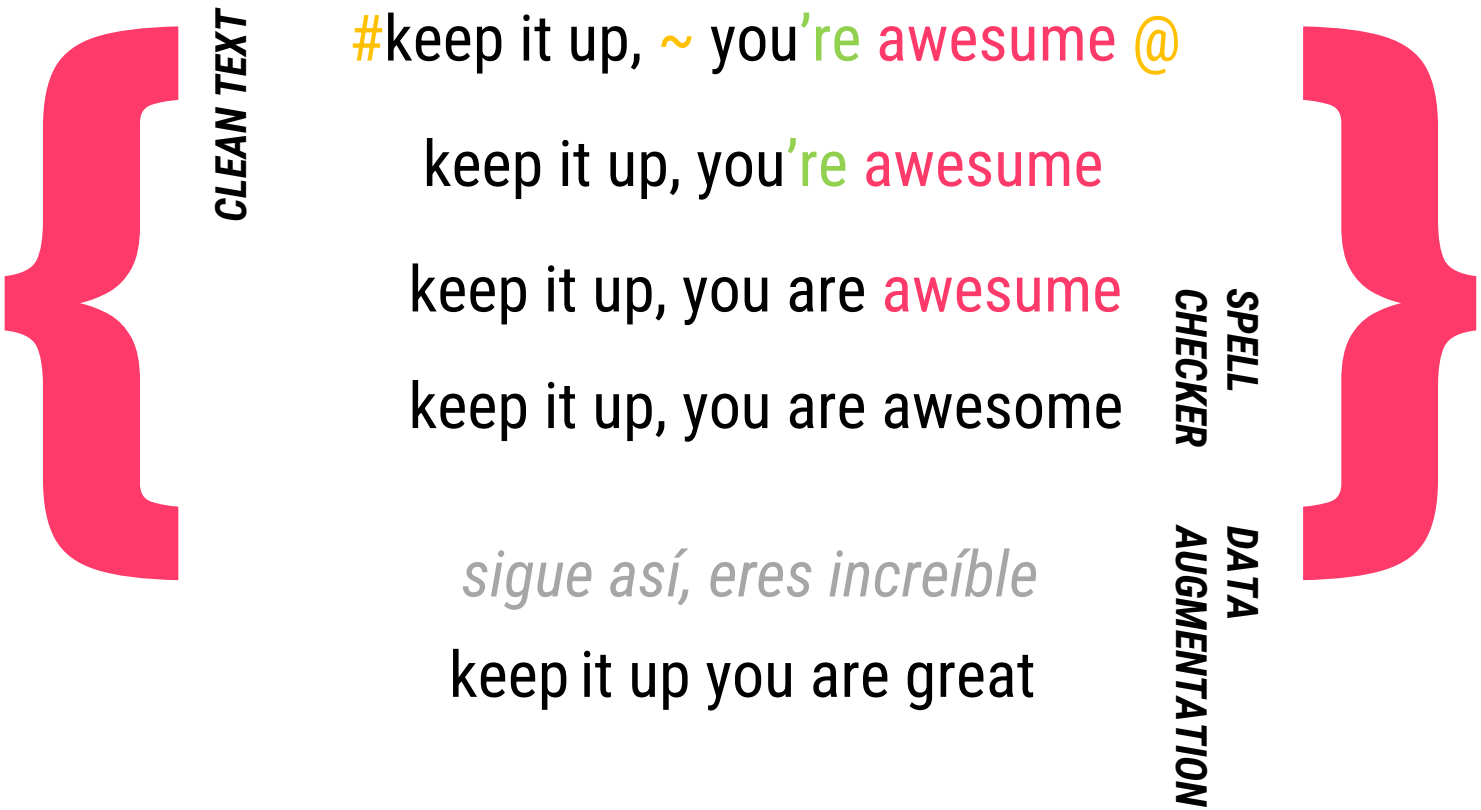


Table of Contents

Table of Contents

Features

Lab Sessions

The chef prepared a delicious pasta dish

Tokens: ['chef', 'prepared', 'delicious', 'pasta', 'dish']

Lemmas: ['chef', 'prepare', 'delicious', 'pasta', 'dish']

Senses / Definitions: ['chef.n.01', 'train.v.02', 'pasta.n.01', 'smasher.n.02']

Synsents

New Features

Semantic Textual Similarity

Ngrams with words and characters (1,2,…5).

Words BiGrams: ['semantic textual', 'textual similarity']

Characters BiGrams: ['se','em','ma',…]

Ngrams with POS Tags.

Bigram Example: ['JJ', 'NNP', 'NNP'] →['JJ NNP', 'NNP NNP']

Features

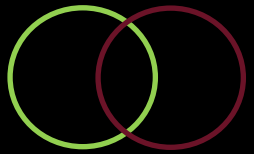
Lab Metrics

Dice **Cosine** **Overlap** **Jaccard**

Similarities based on set relations.

New Metrics

Intersection



Maximum Similarity of Synsents

Count Nº Equal

Longest Common Subsequence

$s1 = \text{BD}$ LCS: BD

$s2 = \text{ABCD}$

Longest Common Substring

$s1 = \text{abcdxyz}$ LCS: abcd

$s2 = \text{xyzabcd}$

Table of Contents

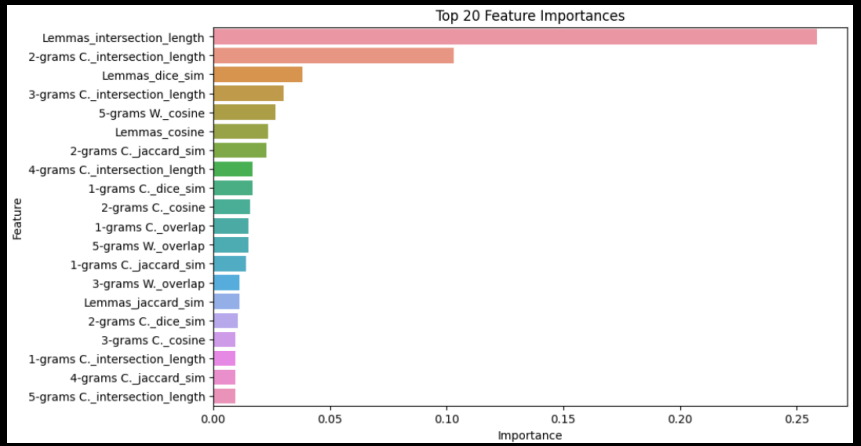
Table of Contents

Model and Results

Lexical	Syntactic	Total
0.7567	0.6572	0.7666

Feature Reduction

Top 20



0.7629

Model and Results

Lexical	Syntactic	Total
0.7567	0.6572	0.7666

PCA → 0.7676

Retains 95% of the variance.

PCA + Data Augmentation → 0.7635

Table of Contents

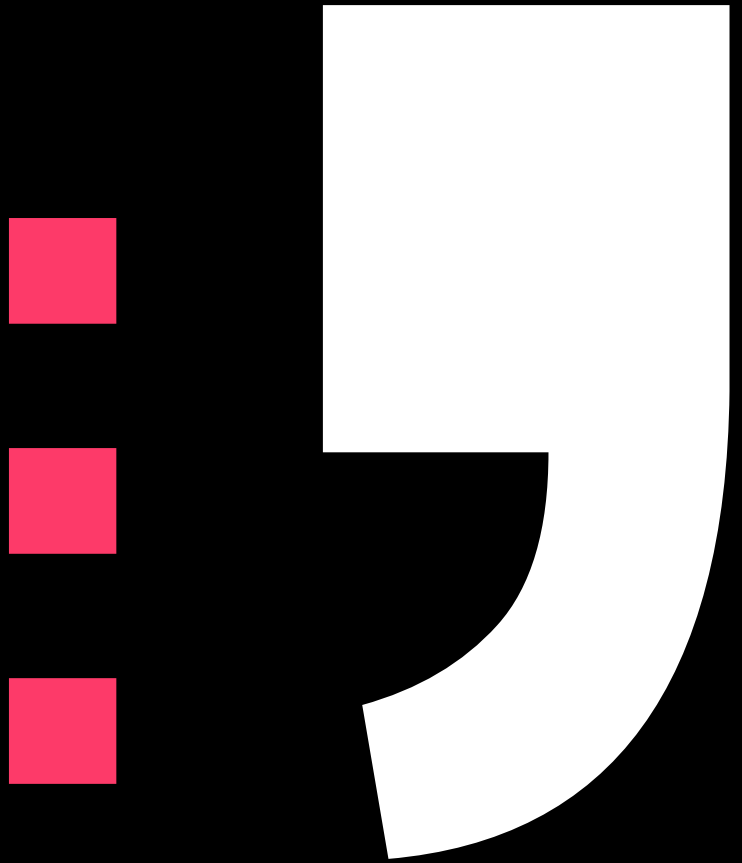
Table of Contents

Conclusions

1. IMPORTANCE OF PRE-PROCESSING

2. GREAT RESULTS THROUGH THE VOTING ENSEMBLE

3. BETTER RESULT BY REDUCING THE DIMENSIONALITY OF THE PCA



Thanks

Do you have any questions?

Unai Gurbindo / Jaume Guasch

BIBLIOGRAPHY

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., & Bašić, B. D. (2012). Takelab: Systems for measuring semantic text similarity. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 441-448).

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 435-440).