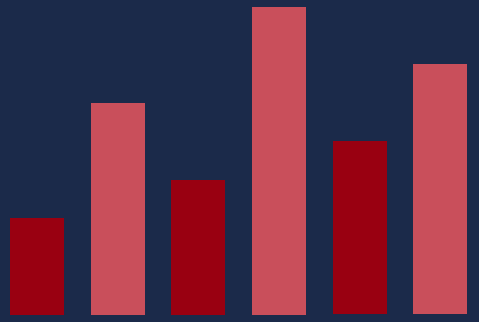
SENTIMENT ANALYSIS • SOCIAL MEDIA • BIG DATA



A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets to News

Sercan Kulcu • Erdogan Dogdu

TOBB University of Economics and Technology, Ankara, Turkey

2016 IEEE Tenth International Conference on Semantic Computing (ICSC) | Laguna Hills, California

Outline

01

Motivation & Problem

Why sentiment analysis of Turkish tweets needs a scalable, language-aware approach

02

System Architecture

Collecting news & tweets, mapping, and distributed sentiment scoring on Hadoop

03

Methods

Zemberek NLP, bag-of-words mapping, Naive Bayes / Complement NB / Logistic Regression

04

Experimental Results

Accuracy, precision, F1, dataset size effects, n-grams, and cluster performance

05

Conclusion

Key findings and future work

Motivation

Understanding public opinion about news requires speed and language sensitivity

1

Real-time need

Public reaction to news must be captured as it happens, requiring a system that scales across distributed infrastructure.

2

Turkish is hard

Agglutinative morphology means countless word forms from a single root — standard English-tuned NLP pipelines underperform.

Research goal: *build a scalable framework that (1) collects Turkish news and tweets, (2) links tweets to the news items they discuss, and (3) classifies each tweet's sentiment — all running on a distributed Hadoop/Mahout pipeline.*

System Architecture

Three-stage pipeline: collect → map → score

Runs continuously in 10-minute batches; results are merged incrementally by a second MapReduce job.

Output per news item

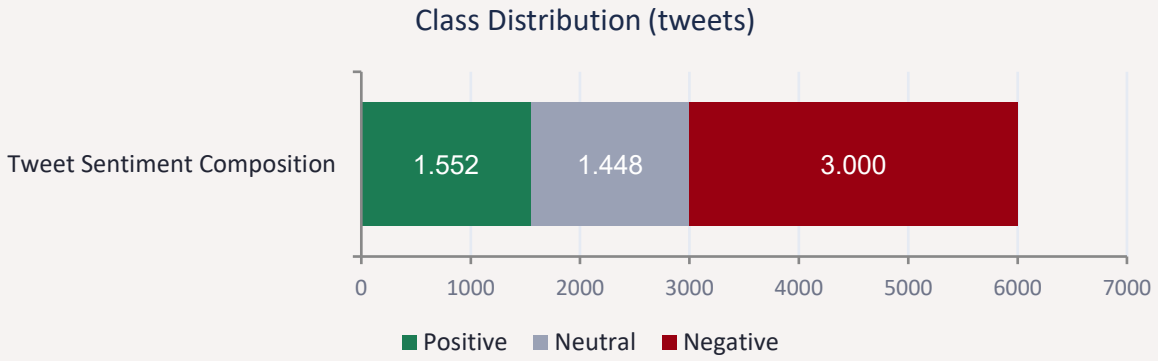
Positive | **Neutral** | **Negative**

Rates how each news article is perceived on Twitter

The Dataset

6,000 manually labeled Turkish tweets, collected in early 2015

6,000	3,000	1,552	1,448
Total tweets	Negative	Positive	Neutral



Dataset characteristics

- Minimum 50 characters per tweet, for meaningful content
- Average length: 9.43 words / 75.3 characters
- Manually labeled and reliability-checked by 10 graduate students
- Split 80% training / 20% test for supervised learning
- Publicly available on GitHub (sercankulcu/twitterdata)

Handling Turkish: The Zemberek NLP Pipeline

Agglutinative morphology requires stemming before any similarity comparison

Stemming Examples

Word	Root	Type
gelmiyorum	gel	verb
gitmedim	git	verb
çok	çok	adjective
erken	erken	adverb
hakaret	hakaret	noun
özgürlük	özgür	noun

Word Similarity Scores

Word 1	Word 2	Score
Yaşasın	yaşasın	0.904
akın	sakın	0.933
kötüyüm	kotuyum	0.768
hakaret	hakret	0.966
akın	yakın	0.933
özgürlük	özgürlüğü	0.935

Similarity threshold: A word pair with similarity score above 0.90 counts as a match. Tweets and news articles are stemmed, stripped of stop words, and compared word-by-word to compute a total similarity score used for tweet-to-news mapping.

Mapping Tweets to News Articles

Bag-of-words similarity scoring (Algorithm 2 in the paper)

1

Parse & clean

Extract words from the tweet and each candidate news article; remove stop words.

2

Stem

Reduce all words to their roots using Zemberek's morphological analyzer.

3

Compare

For every tweet word × news word pair, compute similarity; count matches above 0.90.

4

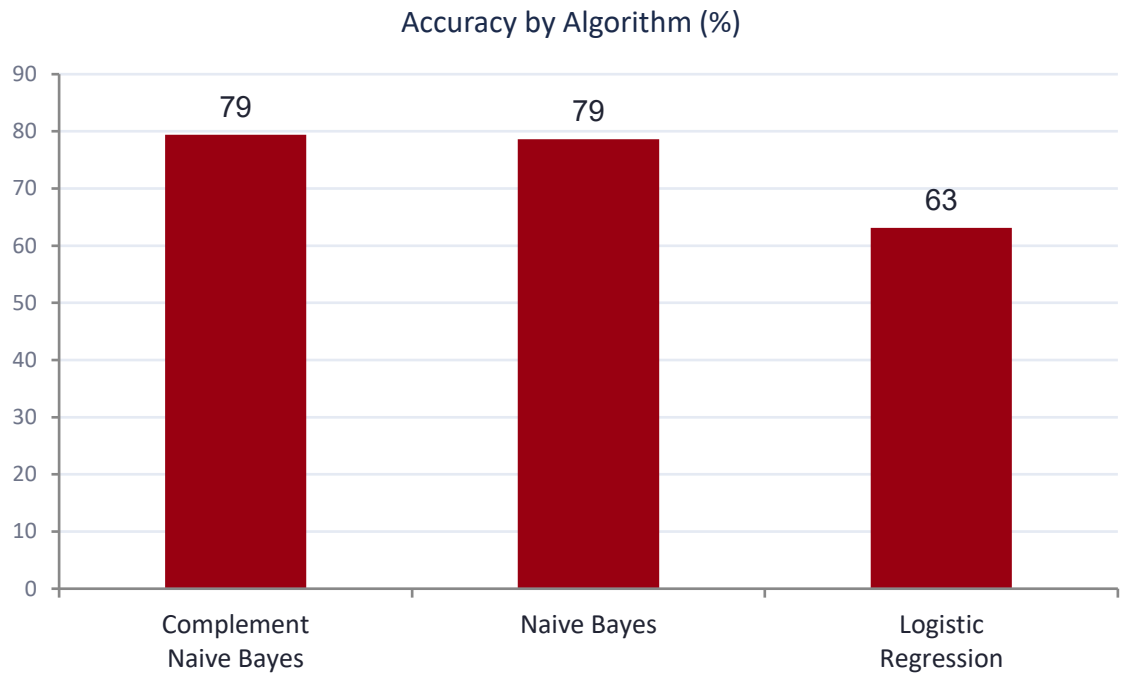
Score & select

Sum matches per news article; return the article with the highest total score.

Result: 218 tweets were mapped to 29 news items and checked manually — 88 correct out of 218 (40.3% accuracy). The authors note this can be improved with lexicon-based strategies.

Classifier Comparison

Naive Bayes, Complement Naive Bayes, and Logistic Regression on the 2-class dataset

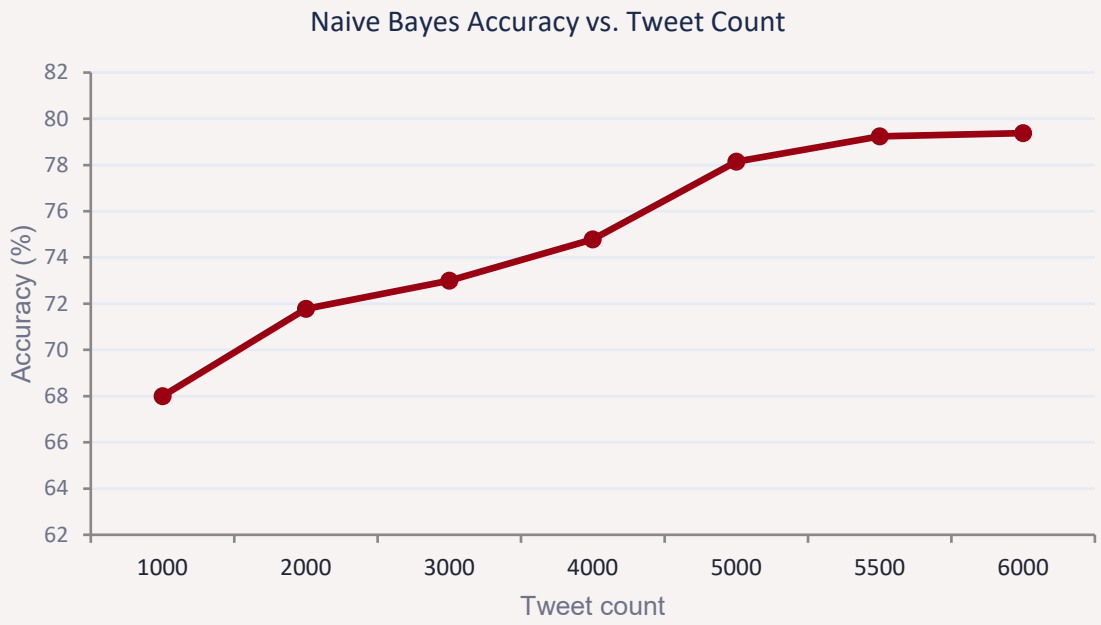


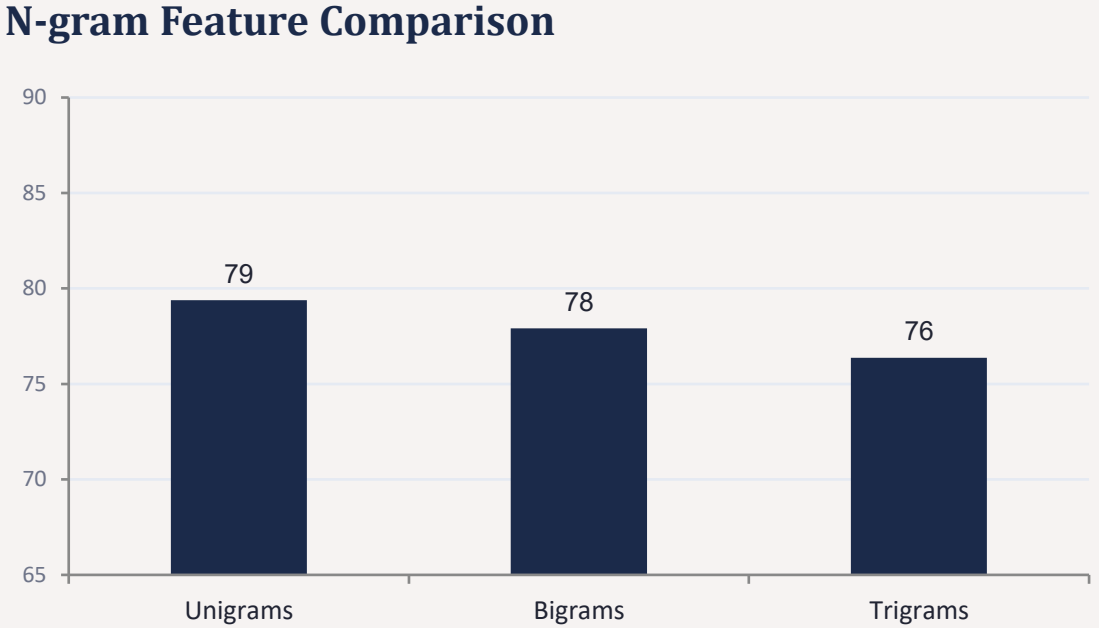
Algorithm	Accuracy (%)	Precision	F1
Complement NB	79.38	0.977	0.880
Naive Bayes	78.61	0.972	0.873
Log. Regression	63.12	0.615	0.620

Complement Naive Bayes corrects for class imbalance (more negative than positive/neutral tweets), giving it the edge over standard Naive Bayes.

What Affects Accuracy?

Dataset size and feature model (n-grams) both matter





Accuracy improves steadily with more training data and plateaus near 6,000 tweets. Unigrams outperform bigrams and trigrams — longer n-grams overfit on this dataset size.

Distributed Execution Performance

Why the system runs on Hadoop MapReduce

Single Node

2,137

seconds

3.4×

faster

Cluster (Six Nodes)

637

seconds

Benchmarked on a 100 MB dataset. The multi-node cluster demonstrates the scalability needed for continuous, real-time sentiment analysis of live tweet streams.

2-Class vs. 3-Class Classification

Including a neutral class makes the task harder

Dataset	Accuracy (%)	Precision	F1
Neg–Pos	79.38	0.977	0.880
Neg–(Pos+Neu)	79.42	0.985	0.880
Neg–Neu–Pos (3-class)	67.62	0.667	0.794
SS-Tweet (3-class)	67.47	0.562	0.593

2-class datasets (higher accuracy)

3-class datasets (lower accuracy)

Learning rate matters too: For Logistic Regression, tuning the learning rate parameter on the 2-class dataset yielded a maximum of only 63.1% accuracy (at rate = 0.9) — still well below Naive Bayes methods. The learning rate is generally chosen by trial and error.

Conclusion

A scalable, Hadoop-based framework collects news and tweets, maps tweets to related news via bag-of-words + Zemberek stemming, and classifies sentiment with distributed MapReduce jobs.

Naive Bayes performs best among the tested classifiers — Complement Naive Bayes reaches 79.38% accuracy, clearly outperforming Logistic Regression (63.12%).

Unigrams and 2-class formulations consistently outperform bigrams/trigrams and 3-class settings on this dataset size.

Future work

Improve tweet-to-news mapping accuracy (currently 40.3%) using lexicon-based strategies, moving beyond simple bag-of-words comparison.

Selected References

[1] Kulcu, S. & Dogdu, E. (2016). A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets to News. IEEE ICSC.

[2] Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

[3] Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC, vol. 10, 1320–1326.

[4] Rennie, J. D. et al. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. ICML, 616–623.

[5] Khuc, V. N. et al. (2012). Towards Building Large-Scale Distributed Systems for Twitter Sentiment Analysis. ACM SAC, 459–464.

[6] Abel, F. et al. (2011). Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. ESWC, 375–389.

Note: Full reference list (20 citations) available in the original paper. Dataset: github.com/sercankulcu/twitterdata