A Proof of Theorem 1

Theorem 1 (Sparse Bernoulli naive Bayes)**.** *Consider the sparse Bernoulli naive Bayes training problem* (SBNB), *with binary data matrix $X \in \{0,1\}^{n\times m}$. The optimal values of the variables are obtained as follows. Set*

$$v := (f^+ + f^-) \circ \log\Big(\frac{f^+ + f^-}{n}\Big) + (n\mathbf{1} - f^+ - f^-) \circ \log\Big(\mathbf{1} - \frac{f^+ + f^-}{n}\Big),$$
$$w := w^+ + w^-, \;\; w^{\pm} := f^{\pm} \circ \log \frac{f^+}{n_{\pm}} + (n_{\pm}\mathbf{1} - f^{\pm}) \circ \log\Big(\mathbf{1} - \frac{f^{\pm}}{n_{\pm}}\Big).$$

Then identify a set $\mathcal{I}$ of indices with the k largest elements in $w - v$, and set θ_*^+, θ_*^- according to

$$\theta^+_{*_i} = \theta^-_{*_i} = \frac{1}{n}(f_i^+ + f_i^-),\ \forall i \notin \mathcal{I}, \quad \theta^{\pm}_{*_i} = \frac{f_i^{\pm}}{n_{\pm}},\ \forall i \in \mathcal{I}.$$

First note that an ℓ_0-norm constraint on a m-vector q can be reformulated as

$$\|q\|_0 \le k \Longleftrightarrow \exists\, \mathcal{I} \subseteq [m],\;\; |\mathcal{I}| \le k \;:\; \forall\, i \notin \mathcal{I}, \quad q_i = 0.$$

Hence problem (SBNB) is equivalent to

$$\max_{\theta^+,\theta^- \in [0,1]^m, \mathcal{I}} \mathcal{L}_{\rm bnb}(\theta^+,\theta^-;X) \;:\; \theta_i^+ = \theta_i^- \;\; \forall i \notin \mathcal{I},\;\; \mathcal{I} \subseteq [m],\;\; |\mathcal{I}| \le k, \tag{18}$$

where the complement of the index set $\mathcal{I}$ encodes the indices where variables θ^+, θ^- agree. Then (18) becomes

$$\begin{aligned} p^* := \max_{\mathcal{I}\subseteq[m],\, |\mathcal{I}|\le k} \quad & \sum_{i\notin\mathcal{I}} \Big(\max_{\theta_i \in [0,1]} (f_i^+ + f_i^-)\log\theta_i + (n - f_i^+ - f_i^-)\log(1-\theta_i)\Big) \\ & + \sum_{i\in\mathcal{I}} \Big(\max_{\theta_i^+ \in [0,1]} f_i^+ \log\theta_i^+ + (n_+ - f_i^+)\log(1-\theta_i^+)\Big) \\ & + \sum_{i\in\mathcal{I}} \Big(\max_{\theta_i^- \in [0,1]} f_i^- \log\theta_i^- + (n_- - f_i^-)\log(1-\theta_i^-)\Big). \end{aligned} \tag{19}$$

where we use the fact that $n_+ + n_- = n$. All the sub-problems in the above can be solved in closed-form, yielding the optimal solutions

$$\theta^+_{*\,i} = \theta^-_{*\,i} = \frac{1}{n}(f_i^+ + f_i^-),\;\; \forall i \notin \mathcal{I}, \quad \text{and} \quad \theta^{\pm}_{*_i} = \frac{f_i^{\pm}}{n_{\pm}},\;\; \forall i \in \mathcal{I}. \tag{20}$$

Plugging the above inside the objective of (18) results in a Boolean formulation, with a Boolean vector u of cardinality $\le k$ such that $\mathbf{1} - u$ encodes indices for which entries of θ^+, θ^- agree:

$$p^* := \max_{u \in \mathcal{C}_k} (\mathbf{1} - u)^\top v + u^\top w,$$

where, for $k \in [m]$:

$$\mathcal{C}_k := \{u \;:\; u \in \{0,1\}^m,\; \mathbf{1}^\top u \le k\},$$

and vectors v, w are as defined in (8):

$$v := (f^+ + f^-) \circ \log\Big(\frac{f^+ + f^-}{n}\Big) + (n\mathbf{1} - f^+ - f^-) \circ \log\Big(\mathbf{1} - \frac{f^+ + f^-}{n}\Big),$$
$$w := w^+ + w^-, \;\; w^{\pm} := f^{\pm} \circ \log \frac{f^+}{n_{\pm}} + (n_{\pm}\mathbf{1} - f^{\pm}) \circ \log\Big(\mathbf{1} - \frac{f^{\pm}}{n_{\pm}}\Big).$$

We obtain

$$p^* = \mathbf{1}^\top v + \max_{u\in\mathcal{C}_k} u^\top (w - v) = \mathbf{1}^\top v + s_k(w - v),$$

where $s_k(\cdot)$ denotes the sum of the k largest elements in its vector argument. Here we have exploited the fact that the map $z := w - v \ge 0$, which in turn implies that

$$s_k(z) = \max_{u\in\{0,1\}^m \,:\, \mathbf{1}^\top u = k} u^\top z = \max_{u\in\mathcal{C}_k} u^\top z.$$

In order to recover an optimal pair (θ_*^+, θ_*^-), we simply identify the set $\mathcal{I}$ of indices with the k largest elements in $w - v$, and set θ_*^+, θ_*^- according to (20).

B Proof of Theorem 2

Theorem 2 (Sparse Multinomial Naive Bayes). *Let $\phi(k)$ be the optimal value of* (SMNB). *Then $\phi(k) \le \psi(k)$, where $\psi(k)$ is the optimal value of the following one-dimensional convex optimization problem*

$$\psi(k) := C + \min_{\alpha\in[0,1]} s_k(h(\alpha)), \tag{USMNB}$$

where C is a constant, $s_k(\cdot)$ is the sum of the top k entries of its vector argument, and for $\alpha \in (0,1)$

$$h(\alpha) := f_+ \circ \log f_+ + f_- \circ \log f_- - (f_+ + f_-) \circ \log(f_+ + f_-) - f_+ \log \alpha - f_- \log(1-\alpha).$$

Further, given an optimal dual variable α_* that solves (USMNB)*, we can reconstruct a primal feasible (sub-optimal) point (θ^+, θ^-) for* (SMNB) *as follows. For α^* optimal for* (USMNB)*, let $\mathcal{I}$ be complement of the set of indices corresponding to the top k entries of $h(\alpha_*)$; then set $B_\pm := \sum_{i\notin\mathcal{I}} f_i^\pm$, and*

$$\theta^+_{*\,i} = \theta^-_{*\,i} = \frac{f_i^+ + f_i^-}{\mathbf{1}^\top(f^+ + f^-)},\ \forall i \in \mathcal{I}, \quad \theta^\pm_{*_i} = \frac{B_+ + B_-}{B_\pm}\frac{f_i^\pm}{\mathbf{1}^\top(f^+ + f^-)},\ \forall i \notin \mathcal{I}. \tag{21}$$

Proof. We begin by deriving the expression for the upper bound $\psi(k)$.

Duality bound. We first derive the bound stated in the theorem. Problem (SMNB) is written

$$(\theta^+_*, \theta^-_*) = \arg \max_{\theta^+,\theta^-\in[0,1]^m} \ f^{+\top}\log\theta^+ + f^{-\top}\log\theta^- \ : \ \begin{array}{l}\mathbf{1}^\top\theta^+ = \mathbf{1}^\top\theta^- = 1,\\ \|\theta^+ - \theta^-\|_0 \le k.\end{array} \tag{SMNB}$$

By weak duality we have $\phi(k) \le \psi(k)$ where

$$\begin{aligned}\psi(k) := \min_{\substack{\mu^+,\mu^-\\ \lambda\ge 0}} \max_{\theta^+,\theta^-\in[0,1]^m} & f^{+\top}\log\theta^+ + f^{-\top}\log\theta^- + \mu^+(1 - \mathbf{1}^\top\theta^+) + \mu^-(1-\mathbf{1}^\top\theta^-)\\ & + \lambda(k - \|\theta^+ - \theta^-\|_0).\end{aligned}$$

The inner maximization is separable across the components of θ^+, θ^- since $\|\theta^+ - \theta^-\|_0 = \sum_{i=1}^m \mathbf{1}_{\{\theta_i^+ \ne \theta_i^-\}}$. To solve it, we thus only need to consider one dimensional problems written

$$\max_{q,r\in[0,1]} f_i^+ \log q + f_i^- \log r - \mu^+ q - \mu^- r - \lambda \mathbb{1}_{\{q\ne r\}}, \tag{22}$$

where $f_i^+, f_i^- > 0$ and $\mu^\pm > 0$ are given. We can split the max into two cases; one case in which $q = r$ and another when $q \ne r$, then compare the objective values of both solutions and take the larger one. Hence (22) becomes

$$\max\Big(\max_{u\in[0,1]} (f_i^+ + f_i^-)\log u - (\mu^+ + \mu^-)u,\ \max_{q,r\in[0,1]} f_i^+\log q + f_i^-\log r - \mu^+ q - \mu^- r - \lambda\Big).$$

Each of the individual maximizations can be solved in closed form, with optimal point

$$u^* = \frac{(f_i^+ + f_i^-)}{\mu^+ + \mu^-}, \quad q^* = \frac{f_i^+}{\mu^+}, \quad r^* = \frac{f_i^-}{\mu^-}. \tag{23}$$

Note that none of u^*, q^*, r^* can be equal to either 0 or 1, which implies $\mu^+, \mu^- > 0$. Hence (22) reduces to

$$\max\left((f_i^+ + f_i^-)\log\left(\frac{(f_i^+ + f_i^-)}{\mu^+ + \mu^-}\right), f_i^+\log\left(\frac{f_i^+}{\mu^+}\right) + f_i^-\log\left(\frac{f_i^-}{\mu^-}\right) - \lambda\right) - (f_i^+ + f_i^-). \tag{24}$$

We obtain, with $S := \mathbf{1}^\top(f^+ + f^-)$,

$$\psi(k) = -S + \min_{\substack{\mu^+,\mu^->0\\ \lambda\ge 0}} \mu^+ + \mu^- + \lambda k + \sum_{i=1}^m \max(v_i(\mu), w_i(\mu) - \lambda). \tag{25}$$

where, for given $\mu = (\mu^+, \mu^-) > 0$,

$$v(\mu) := (f^+ + f^-) \circ \log\left(\frac{f^+ + f^-}{\mu^+ + \mu^-}\right), \quad w(\mu) := f^+ \circ \log\left(\frac{f^+}{\mu^+}\right) + f^- \circ \log\left(\frac{f^-}{\mu^-}\right).$$

Recall the variational form of $s_k(z)$. For a given vector $z \geq 0$, Lemma 11 shows

$$s_k(z) = \min_{\lambda \geq 0} \ \lambda k + \sum_{i=1}^m \max(0, z_i - \lambda).$$

Problem (25) can thus be written

$$\begin{aligned} \psi(k) &= -S + \min_{\substack{\mu > 0 \\ \lambda \geq 0}} \ \mu^+ + \mu^- + \lambda k + \mathbf{1}^\top v(\mu) + \sum_{i=1}^m \max(0, w_i(\mu) - v_i(\mu) - \lambda) \\ &= -S + \min_{\mu > 0} \ \mu^+ + \mu^- + \mathbf{1}^\top v(\mu) + s_k(w(\mu) - v(\mu)), \end{aligned}$$

where the last equality follows from $w(\mu) \geq v(\mu)$, valid for any $\mu > 0$. To prove this, observe that the negative entropy function $x \to x \log x$ is convex, implying that its perspective P also is. The latter is the function with domain $\mathbb{R}_+ \times \mathbb{R}_{++}$, and values for $x \geq 0$, $t > 0$ given by $P(x, t) = x \log(x/t)$. Since P is homogeneous and convex (hence subadditive), we have, for any pair z_+, z_- in the domain of P: $P(z_+ + z_-) \leq P(z_+) + P(z_-)$. Applying this to $z_\pm := (f_i^\pm, \mu_i^+)$ for given $i \in [m]$ results in $w_i(\mu) \geq v_i(\mu)$, as claimed.

We further notice that the map $\mu \to w(\mu) - v(\mu)$ is homogeneous, which motivates the change of variables $\mu_\pm = t\, p_\pm$, where $t = \mu_+ + \mu_- > 0$ and $p_\pm > 0$, $p_+ + p_- = 1$. The problem reads

$$\begin{aligned} \psi(k) &= -S + (f^+ + f^-)^\top \log(f^+ + f^-) + \min_{\substack{t>0,\, p>0, \\ p_+ + p_- = 1}} \{t - S \log t + s_k(H(p))\} \\ &= C + \min_{p > 0,\, p_+ + p_- = 1} \ s_k(H(p)), \end{aligned}$$

where $C := (f^+ + f^-)^\top \log(f^+ + f^-) - S \log S$, because $t = S$ at the optimum, and

$$H(p) := v - f^+ \circ \log p_+ - f^- \circ \log p_-,$$

with

$$v = f^+ \circ \log f^+ + f^- \circ \log f^- - (f^+ + f^-) \circ \log(f^+ + f^-).$$

Solving for $\psi(k)$ thus reduces to a 1D bisection

$$\psi(k) = C + \min_{\alpha \in [0,1]} \ s_k(h(\alpha)),$$

where

$$h(\alpha) := H(\alpha, 1 - \alpha) = v - f^+ \log \alpha - f^- \log(1 - \alpha).$$

This establishes the first part of the theorem. Note that it is straightforward to check that with $k = n$, the bound is exact: $\phi(n) = \psi(n)$.

Primalization. Next we focus on recovering a primal feasible (sub-optimal) point $(\theta^{+\text{sub}}, \theta^{-\text{sub}})$ from the dual bound obtained before. Assume that α_* is optimal for the dual problem (USMNB). We sort the vector $h(\alpha_*)$ and find the indices corresponding to the top k entries. Denote the complement of this set of indices by $\mathcal{I}$. These indices are then the candidates for which $\theta_i^+ = \theta_i^-$ for $i \in \mathcal{I}$ in the primal problem to eliminate the cardinality constraint. Hence we are left with solving

$$\begin{aligned} (\theta^{+\text{sub}}, \theta^{-\text{sub}}) = \arg &\max_{\theta^+, \theta^- \in [0,1]^m} \ f^{+\top} \log \theta^+ + f^{-\top} \log \theta^- \\ &\text{s.t.} \mathbf{1}^\top \theta^+ = \mathbf{1}^\top \theta^- = 1, \\ &\theta_i^+ = \theta_i^-, \ \ i \in \mathcal{I} \end{aligned} \tag{26}$$

or, equivalently

$$\max_{\theta,\theta^+,\theta^-,s\in[0,1]} \sum_{i\in\mathcal{I}}(f_i^+ + f_i^-)\log\theta_i + \sum_{i\notin\mathcal{I}}(f_i^+\log\theta_i^+ + f_i^-\log\theta_i^-) \quad \text{s.t. } \mathbf{1}^\top\theta^+ = \mathbf{1}^\top\theta^- = 1-s, \;\; \mathbf{1}^\top\theta = s. \tag{27}$$

For given $\kappa \in [0,1]$, and $f \in \mathbb{R}^m_{++}$, we have

$$\max_{u\,:\,\mathbf{1}^\top u=\kappa} f^\top \log(u) = f^\top \log f - (\mathbf{1}^\top f)\log(\mathbf{1}^\top f) + (\mathbf{1}^\top f)\log\kappa,$$

with optimal point given by $u^* = (\kappa/(\mathbf{1}^\top f))f$. Applying this to problem (27), we obtain that the optimal value of s is given by

$$s^* = \arg\max_{s\in(0,1)} \{A\log s + B\log(1-s)\} = \frac{A}{A+B},$$

where

$$A := \sum_{i\in\mathcal{I}}(f_i^+ + f_i^-), \;\; B_\pm := \sum_{i\notin\mathcal{I}} f_i^\pm, \;\; B := B_+ + B_- = \mathbf{1}^\top(f^+ + f^-) - A.$$

We obtain

$$\theta_i^{+\text{sub}} = \theta_i^{-\text{sub}} = \frac{s^*}{A}(f_i^+ + f_i^-), \;\; i\in\mathcal{I}, \;\; \theta_i^{\pm\text{sub}} = \frac{(1-s^*)}{B_\pm(A+B)} f_i^\pm, \;\; i\notin\mathcal{I},$$

which further reduces to the expression stated in the theorem. ■

C Proof of Theorem 3

The proof follows from results by (Aubin and Ekeland, 1976) (see also (Ekeland and Temam, 1999; Kerdreux et al., 2017) for a more recent discussion) which are briefly summarized below for the sake of completeness. Given functions f_i, a vector $b \in \mathbb{R}^m$, and vector-valued functions g_i, $i \in [n]$ that take values in $\mathbb{R}^m$, we consider the following problem:

$$\mathrm{h}_P(u) := \min_x \sum_{i=1}^n f_i(x_i) \;\; : \;\; \sum_{i=1}^n g_i(x_i) \le b + u \tag{P}$$

in the variables $x_i \in \mathbb{R}^{d_i}$, with perturbation parameter $u \in \mathbb{R}^m$. We first recall some basic results about conjugate functions and convex envelopes.

Biconjugate and convex envelope. Given a function f, not identically $+\infty$, minorized by an affine function, we write

$$f^*(y) \triangleq \inf_{x\in\mathbf{dom}\,f}\{y^\top x - f(x)\}$$

the conjugate of f, and $f^{**}(y)$ its biconjugate. The biconjugate of f (aka the convex envelope of f) is the pointwise supremum of all affine functions majorized by f (see e.g. (Rockafellar, 1970, Th. 12.1) or (Hiriart-Urruty and Lemaréchal, 1993, Th. X.1.3.5)), a corollary then shows that $\mathbf{epi}(f^{**}) = \overline{\mathbf{Co}(\mathbf{epi}(f))}$. For simplicity, we write $S^{**} = \overline{\mathbf{Co}(S)}$ for any set S in what follows. We will make the following technical assumptions on the functions f_i and g_i in our problem.

Assumption 3. *The functions $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ are proper, 1-coercive, lower semicontinuous and there exists an affine function minorizing them.*

Note that coercivity trivially holds if $\mathbf{dom}(f_i)$ is compact (since f can be set to $+\infty$ outside w.l.o.g.). When Assumption 3 holds, $\mathbf{epi}(f^{**})$, f_i^{**} and hence $\sum_{i=1}^n f_i^{**}(x_i)$ are closed (Hiriart-Urruty and Lemaréchal, 1993, Lem. X.1.5.3). Also, as in e.g. (Ekeland and Temam, 1999), we define the lack of convexity of a function as follows.

Definition 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$, we let*

$$\rho(f) \triangleq \sup_{x\in\mathbf{dom}(f)}\{f(x) - f^{**}(x)\} \tag{28}$$

Many other quantities measure lack of convexity (see e.g. (Aubin and Ekeland, 1976; Bertsekas, 2014) for further examples). In particular, the nonconvexity measure $\rho(f)$ can be rewritten as

$$\rho(f) = \sup_{\substack{x_i \in \mathbf{dom}(f) \\ \mu \in \mathbb{R}^{d+1}}} \left\{ f\left(\sum_{i=1}^{d+1} \mu_i x_i\right) - \sum_{i=1}^{d+1} \mu_i f(x_i) : \mathbf{1}^\top \mu = 1, \mu \geq 0 \right\} \tag{29}$$

when f satisfies Assumption 3 (see (Hiriart-Urruty and Lemaréchal, 1993, Th. X.1.5.4)).

Bounds on the duality gap and the Shapley-Folkman Theorem Let $\mathrm{h}_P(u)^{**}$ be the biconjugate of $\mathrm{h}_P(u)$ defined in (P), then $\mathrm{h}_P(0)^{**}$ is the optimal value of the dual to (P) (Ekeland and Temam, 1999, Lem. 2.3), and (Ekeland and Temam, 1999, Th. I.3) shows the following result.

Theorem 5. *Suppose the functions f_i, g_{ji} in problem* (P) *satisfy Assumption 3 for $i = 1, \ldots, n$, $j = 1, \ldots, m$. Let*

$$\bar{p}_j = (m+1) \max_i \rho(g_{ji}), \quad \text{for } j = 1, \ldots, m \tag{30}$$

then

$$\mathrm{h}_P(\bar{p}) \leq \mathrm{h}_P(0)^{**} + (m+1) \max_i \rho(f_i). \tag{31}$$

where $\rho(\cdot)$ is defined in Def. 4.

We are now ready to prove Theorem 3, whose proof follows from Theorem 5 above.

Theorem 6 (Quality of Sparse Multinomial Naive Bayes Relaxation)**.** *Let $\phi(k)$ be the optimal value of* (SMNB) *and $\psi(k)$ that of the convex relaxation in* (USMNB)*, we have for $k \geq 4$,*

$$\psi(k-4) \leq \phi(k) \leq \psi(k) \leq \phi(k+4).$$

for $k \geq 4$.

Proof. Problem (SMNB) is *separable* and can be written in perturbation form as in the result by (Ekeland and Temam, 1999, Th. I.3) recalled in Theorem 5, to get

$$\begin{array}{rll} \mathrm{h}_P(u) = & \min_{q,r} & -f^{+\top} \log q - f^{-\top} \log r \\ & \text{subject to} & \mathbf{1}^\top q = 1 + u_1, \\ & & \mathbf{1}^\top r = 1 + u_2, \\ & & \sum_{i=1}^m \mathbf{1}_{q_i \neq r_i} \leq k + u_3 \end{array} \tag{32}$$

in the variables $q, r \in [0,1]^m$, where $u \in \mathbb{R}^3$ is a perturbation vector. By construction, we have $\phi(k) = -\mathrm{h}_P(0)$ and $\phi(k+l) = -\mathrm{h}_P((0,0,l))$. Note that the functions $\mathbf{1}_{q_i \neq r_i}$ are lower semicontinuous and, because the domain of problem (SMNB) is compact, the functions

$$f_i^+ \log q_i + q_i + f_i^- \log r_i + r_i + \mathbf{1}_{q_i \neq r_i}$$

are 1-coercive for $i = 1, \ldots, m$ on the domain and satisfy Assumption 3 above.

Now, because $q, r \geq 0$ with $\mathbf{1}^\top q = \mathbf{1}^\top r = 1$, we have $q - r \in [-1,1]^m$ and the convex envelope of $\mathbf{1}_{q_i \neq r_i}$ on $q, r \in [0,1]^m$ is $|q_i - r_i|$, hence the *lack of convexity* (29) of $\mathbf{1}_{q_i \neq r_i}$ on $[0,1]^2$ is bounded by one, because

$$\rho(\mathbf{1}_{x \neq y}) := \sup_{x,y \in [0,1]} \{\mathbf{1}_{y \neq x} - |x - y|\} = 1$$

which means that $\max_{i=1,\ldots,n} \rho(g_{3i}) = 1$ in the statement of Theorem 5. The fact that the first two constraints in problem (32) are convex means that $\max_{i=1,\ldots,n} \rho(g_{ji}) = 0$ for $j = 1, 2$, and the perturbation vector in (30) is given by $\bar{p} = (0,0,4)$, because there are three constraints in problem (32) so $m = 3$ in (30), hence

$$\mathrm{h}_P(\bar{p}) = \mathrm{h}_P((0,0,4)) = -\phi(k+4).$$

The objective function being convex separable, we have $\max_{i=1,\ldots,n} \rho(f_i) = 0$. Theorem 5 then states that

$$\mathrm{h}_P(\bar{p}) = \mathrm{h}_P((0,0,4)) = -\phi(k+4) \leq \mathrm{h}_P(0)^{**} + 0 = -\psi(k)$$

because $-\mathrm{h}_P(0)^{**}$ is the optimal value of the dual to $\phi(k)$ which is here $\psi(k)$ defined in Theorem 2. The other bound in (15), namely $\phi(k) \leq \psi(k)$, follows directly from weak duality. ∎

Primalization. We first derive the second dual of problem (P), i.e. the dual of problem (USMNB), which will be used to extract good primal solutions.

Proposition 7. *A dual of problem* (USMNB) *is written*

$$\begin{array}{ll} \textit{max.} & z^\top (g \circ \log(g)) + x^\top (f^+ \circ \log(f^+) + f^- \circ \log(f^-)) + (x^\top g)\log(x^\top g) - (x^\top g) \\ & -(\mathbf{1}^\top g)\log(\mathbf{1}^\top g) - (x^\top f^+)\log(x^\top f^+) - (x^\top f^-)\log(x^\top f^-) \\ \textit{s.t.} & x + z = \mathbf{1},\ \ \mathbf{1}^\top x \le k,\ \ x \ge 0,\ \ z \ge 0 \end{array} \tag{D}$$

in the variables $x, z \in \mathbb{R}^n$. Furthermore, strong duality holds between the dual (USMNB) *and its dual* (D)*.*

Proof. The dual optimum value $\psi(k)$ in (USMNB) can be written as in (25),

$$\psi(k) = -S + \min_{\substack{\mu^+,\mu^- > 0 \\ \lambda \ge 0}} \ \mu^+ + \mu^- + \lambda k + \sum_{i=1}^m \max(v_i(\mu), w_i(\mu) - \lambda).$$

with $S := \mathbf{1}^\top (f^+ + f^-)$, and

$$v(\mu) := (f^+ + f^-) \circ \log\left(\frac{f^+ + f^-}{\mu^+ + \mu^-}\right), \ \ w(\mu) := f^+ \circ \log\left(\frac{f^+}{\mu^+}\right) + f^- \circ \log\left(\frac{f^-}{\mu^-}\right).$$

for given $\mu = (\mu^+, \mu^-) > 0$. This can be rewritten

$$\min_{\substack{\mu^+,\mu^- > 0 \\ \lambda \ge 0}} \ \max_{\substack{x+z=1 \\ x,z \ge 0}} \ \mu^+ + \mu^- - S + \lambda(k - \mathbf{1}^\top x) + z^\top v(\mu) + x^\top w(\mu)$$

using additional variables $x, z \in \mathbb{R}^n$, or again

$$\min_{\substack{\mu^+,\mu^- > 0 \\ \lambda \ge 0}} \ \max_{\substack{x+z=1 \\ x,z \ge 0}} \ \begin{array}{l} \lambda(k - \mathbf{1}^\top x) - (x+z)^\top g - (z^\top g)\log(\mu^+ + \mu^-) + z^\top (g \circ \log(g)) \\ -(x^\top f^+)\log(\mu^+) - (x^\top f^-)\log(\mu^-) \\ +x^\top (f^+ \circ \log(f^+) + f^- \circ \log(f^-)) + \mu^+ + \mu^- \end{array} \tag{33}$$

calling $g = f^+ + f^-$. Strong duality holds in this min max problem so we can switch the min and the max. Writing $\mu_\pm = t\,p_\pm$, where $t = \mu_+ + \mu_-$ and $p_\pm > 0$, $p^+ + p^- = 1$ the Lagrangian becomes

$$\begin{array}{rcl} L(p_+, p_-, t, \lambda, x, z, \alpha) & = & \mathbf{1}^\top \nu - z^\top \nu - x^\top \nu + \lambda k - \lambda \mathbf{1}^\top x - \mathbf{1}^\top g - (z^\top g)\log(t) \\ & & -(x^\top f^+)\log(t\,p_+) - (x^\top f^-)\log(t\,p_-) + t \\ & & +z^\top (g \circ \log(g)) + x^\top (f^+ \circ \log(f^+) + f^- \circ \log(f^-)) \\ & & +\alpha(p_+ + p_- - 1), \end{array}$$

where α is the dual variable associated with the constraint $p_+ + p_- = 1$. The dual of problem (USMNB) is then written

$$\sup_{\{x \ge 0, z \ge 0, \alpha\}} \ \inf_{\substack{p_+ \ge 0, p_- \ge 0, \\ t \ge 0, \lambda \ge 0}} L(p_+, p_-, t, \mu^-, \lambda, x, z, \alpha)$$

The inner infimum will be $-\infty$ unless $\mathbf{1}^\top x \le k$, so the dual becomes

$$\sup_{\substack{x+z=\mathbf{1}, \mathbf{1}^\top x \le k, \\ x \ge 0, z \ge 0, \alpha}} \ \inf_{\substack{p_+ \ge 0, p_- \ge 0, \\ t \ge 0}} \ \begin{array}{l} z^\top (g \circ \log(g)) + x^\top (f^+ \circ \log(f^+) + f^- \circ \log(f^-)) \\ -(x^\top f^+)(\log t + \log(p_+)) - (x^\top f^-)(\log t + \log(p_-)) \\ +t - \mathbf{1}^\top g - (z^\top g)\log(t) + \alpha(p_+ + p_- - 1) \end{array}$$

and the first order optimality conditions in t, p_+, p_- yield

$$\begin{array}{rcl} t & = & \mathbf{1}^\top g \\ p_+ & = & (x^\top f^+)/\alpha \\ p_- & = & (x^\top f^-)/\alpha \end{array} \tag{34}$$

which means the above problem reduces to

$$\begin{array}{ll}\sup\limits_{\substack{x+z=\mathbf{1},\mathbf{1}^\top x\leq k,\\ x\geq 0, z\geq 0,\alpha}} & z^\top(g\circ\log(g))+x^\top(f^+\circ\log(f^+)+f^-\circ\log(f^-))\\ & -(\mathbf{1}^\top g)\log(\mathbf{1}^\top g)-(x^\top f^+)\log(x^\top f^+)-(x^\top f^-)\log(x^\top f^-)\\ & +(x^\top g)\log\alpha-\alpha\end{array}$$

and setting in $\alpha = x^\top g$ leads to the dual in (D). ■

We now use this last result to better characterize scenarios where the bound produced by problem (USMNB) is tight and recovers an optimal solution to problem (SMNB).

Proposition 8. *Given $k>0$, let $\phi(k)$ be the optimal value of* (SMNB)*. Given an optimal solution (x,z) of problem* (D)*, let $J=\{i : x_i\notin\{0,1\}\}$ be the set of indices where x_i, z_i are not binary in $\{0,1\}$. There is a feasible point $\bar\theta,\bar\theta^+,\bar\theta^-$ of problem* (SMNB) *for $\bar k = k+|J|$, with objective value OPT such that*

$$\phi(k)\leq OPT\leq \phi(k+|J|).$$

Proof. Using the fact that

$$\max_x\ a\log(x)-bx = a\log\left(\frac{a}{b}\right)-a$$

the max min problem in (33) can be rewritten as

$$\begin{array}{llll}\max\limits_{\substack{x+z=1\\ x,z\geq 0}} & \min\limits_{\substack{\mu^+,\mu^->0\\ \lambda\geq 0}} & \max\limits_{\theta,\theta^+,\theta^-} & \lambda(k-\mathbf{1}^\top x)+z^\top(g\circ\log\theta)\\ & & & +x^\top(f^+\circ\log\theta^+)+x^\top(f^-\circ\log\theta^-)\\ & & & +\mu^+(1-z^\top\theta-x^\top\theta^+)+\mu^-(1-z^\top\theta-x^\top\theta^-)\end{array}\tag{35}$$

in the additional variables $\theta,\theta^+,\theta^-\in\mathbb{R}^n$, with (23) showing that

$$\theta_i=\frac{(f_i^++f_i^-)}{\mu^++\mu^-},\quad \theta_i^+=\frac{f_i^+}{\mu^+},\quad \theta_i^-=\frac{f_i^-}{\mu^-}.$$

at the optimum. Strong duality holds in the inner min max, which means we can also rewrite problem (D) as

$$\max_{\substack{x+z=1\\ x,z\geq 0}}\ \max_{\substack{z^\top\theta+x^\top\theta^+\leq 1\\ z^\top\theta+x^\top\theta^-\leq 1\\ x^\top\mathbf{1}\leq k}} z^\top(g\circ\log\theta)+x^\top(f^+\circ\log\theta^++f^-\circ\log\theta^-)\tag{36}$$

or again, in epigraph form

$$\begin{array}{ll}\text{max.} & r\\ \text{s.t.} & \begin{pmatrix} r\\1\\1\\k\end{pmatrix}\in\begin{pmatrix}0\\ \mathbb{R}_+\\ \mathbb{R}_+\\ \mathbb{R}_+\end{pmatrix}+\sum_{i=1}^n\left\{z_i\begin{pmatrix}g_i\log\theta_i\\ \theta_i\\ \theta_i\\ 0\end{pmatrix}+x_i\begin{pmatrix}f_i^+\log\theta_i^++f_i^-\log\theta_i^-\\ \theta_i^+\\ \theta_i^-\\ 1\end{pmatrix}\right\}\end{array}\tag{37}$$

Suppose the optimal solutions $x^\star, z^\star$ of problem (D) are binary in $\{0,1\}^n$ and let $\mathcal{I}=\{i : z_i=0\}$, then problem (hence problem (D)) reads

$$\begin{array}{c}(\theta^{+\text{sub}},\theta^{-\text{sub}})=\arg\max\limits_{\theta^+,\theta^-\in[0,1]^m} f^{+\top}\log\theta^++f^{-\top}\log\theta^-\\ \text{s.t.}\mathbf{1}^\top\theta^+=\mathbf{1}^\top\theta^-=1,\\ \theta_i^+=\theta_i^-,\ \ i\in\mathcal{I}.\end{array}\tag{38}$$

which is exactly (38). This means that the optimal values of problem (38) and (D) are equal, so that the relaxation is tight and $\theta_i^+=\theta_i^-$ for $i\in\mathcal{I}$. Suppose now that some coefficients x_i are not binary. Let us call J the set $J=\{i : x_i\notin\{0,1\}\}$. As in (Ekeland and Temam, 1999, Th. I.3), we define new solutions $\bar\theta,\bar\theta^+,\bar\theta^-$ and $\bar x,\bar z$ as follows,

$$\left\{\begin{array}{ll}\bar\theta_i=\theta_i,\ \bar\theta_i^+=\theta_i^+,\ \bar\theta_i^-=\theta_i^- \text{ and } \bar z_i=z_i,\ \bar x_i=x_i & \text{if } i\notin J\\ \bar\theta_i=0,\ \bar\theta_i^+=z_i\theta+x_i\theta_i^+,\ \bar\theta_i^-=z_i\theta+x_i\theta_i^- \text{ and } \bar z_i=0,\ \bar x_i=1 & \text{if } i\in J\end{array}\right.$$

By construction, the points $\bar\theta, \bar\theta^+, \bar\theta^-$ and $\bar z, \bar x$ satisfy the constraints $\bar z^\top \bar\theta + \bar x^\top \bar\theta^+ \leq 1$, $\bar z^\top \bar\theta + \bar x^\top \bar\theta^- \leq 1$ and $\bar x^\top \mathbf{1} \leq k$. We also have $\bar x^\top \leq k + |J|$ and

$$\begin{array}{rl} & z^\top((f^+ + f^-)\circ \log\theta) + x^\top(f^+ \circ \log\theta^+ + f^- \circ \log\theta^-) \\ \leq & \bar z^\top((f^+ + f^-)\circ \log\bar\theta) + \bar x^\top(f^+ \circ \log\bar\theta^+ + f^- \circ \log\bar\theta^-) \end{array}$$

by concavity of the objective, hence the last inequality. ■

We will now use the Shapley-Folkman theorem to bound the number of nonbinary coefficients in Proposition 7 and construct a solution to (D) satisfying the bound in Theorem 3.

Proposition 9. *There is a solution to problem* (D) *with at most four nonbinary pairs* (x_i, z_i)*.*

Proof. Suppose $(x^\star, z^\star, r^\star)$ and $(\theta, \theta_i^+, \theta_i^-)$ solve problem (D) written as in (C), we get

$$\begin{pmatrix} r^\star \\ 1-s_1 \\ 1-s_2 \\ k-s_3 \end{pmatrix} = \sum_{i=1}^n \left\{ z_i \begin{pmatrix} g_i \log\theta_i \\ \theta_i \\ \theta_i \\ 0 \end{pmatrix} + x_i \begin{pmatrix} f_i^+ \log\theta_i^+ + f_i^- \log\theta_i^- \\ \theta_i^+ \\ \theta_i^- \\ 1 \end{pmatrix} \right\} \tag{39}$$

where $s_1, s_2, s_3 \geq 0$. This means that the point $(r^\star, 1-s_1, 1-s_1, k-s_3)$ belongs to a Minkowski sum of segments, with

$$\begin{pmatrix} r^\star \\ 1-s_1 \\ 1-s_2 \\ k-s_3 \end{pmatrix} \in \sum_{i=1}^n \mathbf{Co}\left(\left\{ \begin{pmatrix} g_i \log\theta_i \\ \theta_i \\ \theta_i \\ 0 \end{pmatrix}, \begin{pmatrix} f_i^+ \log\theta_i^+ + f_i^- \log\theta_i^- \\ \theta_i^+ \\ \theta_i^- \\ 1 \end{pmatrix} \right\} \right) \tag{40}$$

The Shapley-Folkman theorem (Starr, 1969) then shows that

$$\begin{array}{rl} \begin{pmatrix} r^\star \\ 1-s_1 \\ 1-s_2 \\ k-s_3 \end{pmatrix} \in & \displaystyle\sum_{[1,n]\setminus\mathcal{S}} \left\{ \begin{pmatrix} g_i \log\theta_i \\ \theta_i \\ \theta_i \\ 0 \end{pmatrix}, \begin{pmatrix} f_i^+ \log\theta_i^+ + f_i^- \log\theta_i^- \\ \theta_i^+ \\ \theta_i^- \\ 1 \end{pmatrix} \right\} \\ & \displaystyle + \sum_{\mathcal{S}} \mathbf{Co}\left(\left\{ \begin{pmatrix} g_i \log\theta_i \\ \theta_i \\ \theta_i \\ 0 \end{pmatrix}, \begin{pmatrix} f_i^+ \log\theta_i^+ + f_i^- \log\theta_i^- \\ \theta_i^+ \\ \theta_i^- \\ 1 \end{pmatrix} \right\} \right) \end{array}$$

where $|\mathcal{S}| \leq 4$, which means that there exists a solution to (D) with at most four nonbinary pairs (x_i, z_i) with indices $i \in \mathcal{S}$. ■

In our case, since the Minkowski sum in (40) is a polytope (as a Minkowski sum of segments), the Shapley-Folkman result reduces to a direct application of the fundamental theorem of linear programming, which allows us to reconstruct the solution of Proposition 9 by solving a linear program.

Proposition 10. *Given* $(x^\star, z^\star, r^\star)$ *and* $(\theta, \theta_i^+, \theta_i^-)$ *solving problem* (D)*, we can reconstruct a solution* (x, z) *solving problem* (7)*, such that at most four pairs* (x_i, z_i) *are nonbinary, by solving*

$$\begin{array}{ll} \textit{min.} & c^\top x \\ \textit{s.t.} & \sum_{i=1}^n (1-x_i) g_i \log\theta_i + x_i(f_i^+ \log\theta_i^+ + f_i^- \log\theta_i^-) = r^\star \\ & \sum_{i=1}^n (1-x_i)\theta_i + x_i\theta_i^+ \leq 1 \\ & \sum_{i=1}^n (1-x_i)\theta_i + x_i\theta_i^- \leq 1 \\ & \sum_{i=1}^n x_i \leq k \\ & 0 \leq x \leq 1 \end{array} \tag{41}$$

which is a linear program in the variable $x \in \mathbb{R}^n$ *where* $c \in \mathbb{R}^n$ *is e.g. a i.i.d. Gaussian vector.*

Proof. Given $(x^\star, z^\star, r^\star)$ and $(\theta, \theta_i^+, \theta_i^-)$ solving problem (D), we can reconstruct a solution (x, z) solving problem (7), by solving (41) which is a linear program in the variable $x \in \mathbb{R}^n$ where $c \in \mathbb{R}^n$ is e.g. a i.i.d. Gaussian vector. This program has $2n+4$ constraints, at least n of which will be saturated at the optimum. In particular, at least $n-4$ constraints in $0 \leq x \leq 1$ will be saturated so at least $n-4$ coefficients x_i will be binary at the optimum, idem for the corresponding coefficients $z_i = 1 - x_i$. ■

Proposition 10 shows that solving the linear program in (41) as a postprocessing step will produce a solution to problem (D) with at most $n-4$ nonbinary coefficient pairs (x_i, z_i). Proposition 8 then shows that this solution satisfies

$$\phi(k) \leq OPT \leq \phi(k+4).$$

which is the bound in Theorem (3).

Finally, we show a technical lemma linking the dual solution (x, z) in (D) above and the support of the k largest coefficients in the computation of $s_k(h(\alpha))$ in theorem 2.

Lemma 11. *Given $c \in \mathbb{R}_+^n$, we have*

$$s_k(c) = \min_{\lambda \geq 0} \lambda k + \sum_{i=1}^n \max(0, c_i - \lambda) \tag{42}$$

and given k, $\lambda \in [c_{[k+1]}, c_{[k]}]$ at the optimum, where $c_{[1]} \geq \ldots \geq c_{[n]}$. Its dual is written

$$\begin{array}{ll} \textit{max.} & x^\top c \\ \textit{s.t.} & \mathbf{1}^\top x \leq k \\ & x + z = 1 \\ & 0 \leq z, x \end{array} \tag{43}$$

When all coefficients c_i are distinct, the optimum solutions x, z of the dual have at most one nonbinary coefficient each, i.e. $x_i, z_i \in (0,1)$ for a single $i \in [1, n]$. If in addition $c_{[k]} > 0$, the solution to (43) is binary.

Proof. Problem (42) can be written

$$\begin{array}{ll} \text{min.} & \lambda k + \mathbf{1}^\top t \\ \text{s.t.} & c - \lambda \mathbf{1} \leq t \\ & 0 \leq t \end{array}$$

and its Lagrangian is then

$$L(\lambda, t, z, x) = \lambda k + \mathbf{1}^\top t + x^\top (c - \lambda \mathbf{1} - t) + z^\top t.$$

The dual to the minimization problem (42) reads

$$\begin{array}{ll} \text{max.} & x^\top c \\ \text{s.t.} & \mathbf{1}^\top x \leq k \\ & x + z = 1 \\ & 0 \leq z, x \end{array}$$

in the variable $w \in \mathbb{R}^n$, its optimum value is $s_k(z)$. By construction, given $\lambda \in [c_{[k+1]}, c_{[k]}]$, only the k largest terms in $\sum_{i=1}^m \max(0, c_i - \lambda)$ are nonzero, and they sum to $s_k(c) - k\lambda$. The KKT optimality conditions impose

$$x_i(c_i - \lambda - t_i) = 0 \quad \text{and} \quad z_i t_i = 0, \quad i = 1, \ldots, n$$

at the optimum. This, together with $x + z = 1$ and $t, x, z \geq 0$, means in particular that

$$\begin{cases} x_i = 0, z_i = 1, & \text{if } c_i - \lambda < 0 \\ x_i = 0, z_i = 1, \text{ or } x_i = 1, z_i = 0 & \text{if } c_i - \lambda > 0 \end{cases} \tag{44}$$

the result of the second line comes from the fact that if $c_i - \lambda > 0$ and $t_i = c_i - \lambda$ then $z_i = 0$ hence $x_i = 1$, if on the other hand $t_i \neq c_i - \lambda$, then $x_i = 0$ hence $z_i = 1$. When the coefficients c_i are all distinct, $c_i - \lambda = 0$ for at most a single index i and (44) yields the desired result. When $c_{[k]} > 0$ and the c_i are all distinct, then the only way to enforce zero gap, i.e.

$$x^\top c = s_k(c)$$

is to set the corresponding coefficients of x_i to one. ■

D Details on Datasets

This section details the data sets used in our experiments.

Downloading data sets.

1. AMZN The complete Amazon reviews data set was collected from here; only a subset of this data was used which can be found here. This data set was randomly split into 80/20 train/test.
2. IMDB The large movie review (or IMDB) data set was collected from here and was already split 50/50 into train/test.
3. TWTR The Twitter Sentiment140 data set was downloaded from here and was pre-processed according to the method highlighted here.
4. MPQA The MPQA opinion corpus can be found here and was pre-processed using the code found here.
5. SST2 The Stanford Sentiment Treebank data set was downloaded from here and the pre-processing code can be found here.

Creating feature vectors. After all data sets were downloaded and pre-processed, the diffeent types of feature vectors were constructed using `CounterVectorizer` and `TfidfVectorizer` from Sklearn (Pedregosa et al., 2011). Counter vector, tf-idf, and tf-idf word bigrams use the `analyzer = 'word'` specification while the tf-idf char bigrams use `analyzer = 'char'`.

Two-stage procedures. For experiments 2 and 3, all standard models were trained in Sklearn (Pedregosa et al., 2011). In particular, the following settings were used in stage 2 for each model

1. `LogisticRegression(penalty='l2', solver='lbfgs', C =1e4, max_iter=1e2)`
2. `LinearSVC(C = 1e4)`
3. `MultinomialNB(alpha=a)`

In the first stage of the two stage procedures, the following settings were used for each of the different feature selection methods

1. `LogisticRegression(random_state=0, C =` λ_1`,penalty='l1',solver='saga', max_iter=1e2)`
2. `clf = LogisticRegression(C = 1e4, penalty='l2',  solver = 'lbfgs', max_iter = 1e2).fit(train_x,train_y)`
   `selector_log = RFE(clf,` k`), step=0.3)`
3. `Lasso(alpha =` λ_2`, selection='cyclic', tol = 1e-5)`
4. `LinearSVC(C =`λ_3`,  penalty='l1',dual=False)`
5. `clf = LinearSVC(C = 1e4, penalty='l2',dual=False).fit(train_x,train_y)`
   `selector_svm = RFE(clf,`k`, step=0.3)`
6. `MultinomialNB(alpha=a)`

where λ_i are hyper-parameters used by the ℓ_1 methods to achieve a desired sparsity level k. a is a hyper-parameter for the different MNB models which we compute using cross validation (explained below).

Hyper-parameters. For each of the ℓ_1 methods we manually do a grid search over all hyper-parameters to achieve an approximate desired sparsity pattern. For determining the hyper-parameter for the MNB models, we employ 10-fold cross validation on each data set for each type of feature vector and determine the best value of a. In total, this is $16 + 20 = 36$ values of a – 16 for experiment 2 and 20 for experiment 3. In experiment 2, we do not use the twitter data set since computing the λ_i's to achieve a desired sparsity pattern for the ℓ_1 based feature selection methods was computationally intractable.

Experiment 2 and 3: full results. Here we show the results of experiments 2 and 3 for all the data sets. All error bars represents 10 separate simulations where each simulation is a different appropriately-sized train-test split (as per Table 1). As seen in Figure 1, the SVM-ℓ_1 model was unable to converge and hence has an accuracy of 50%. This was in spite of manually adjusting `max_iter=1e7` and using the liblinear solver which is default for LinearSVC in sci-kit learn.

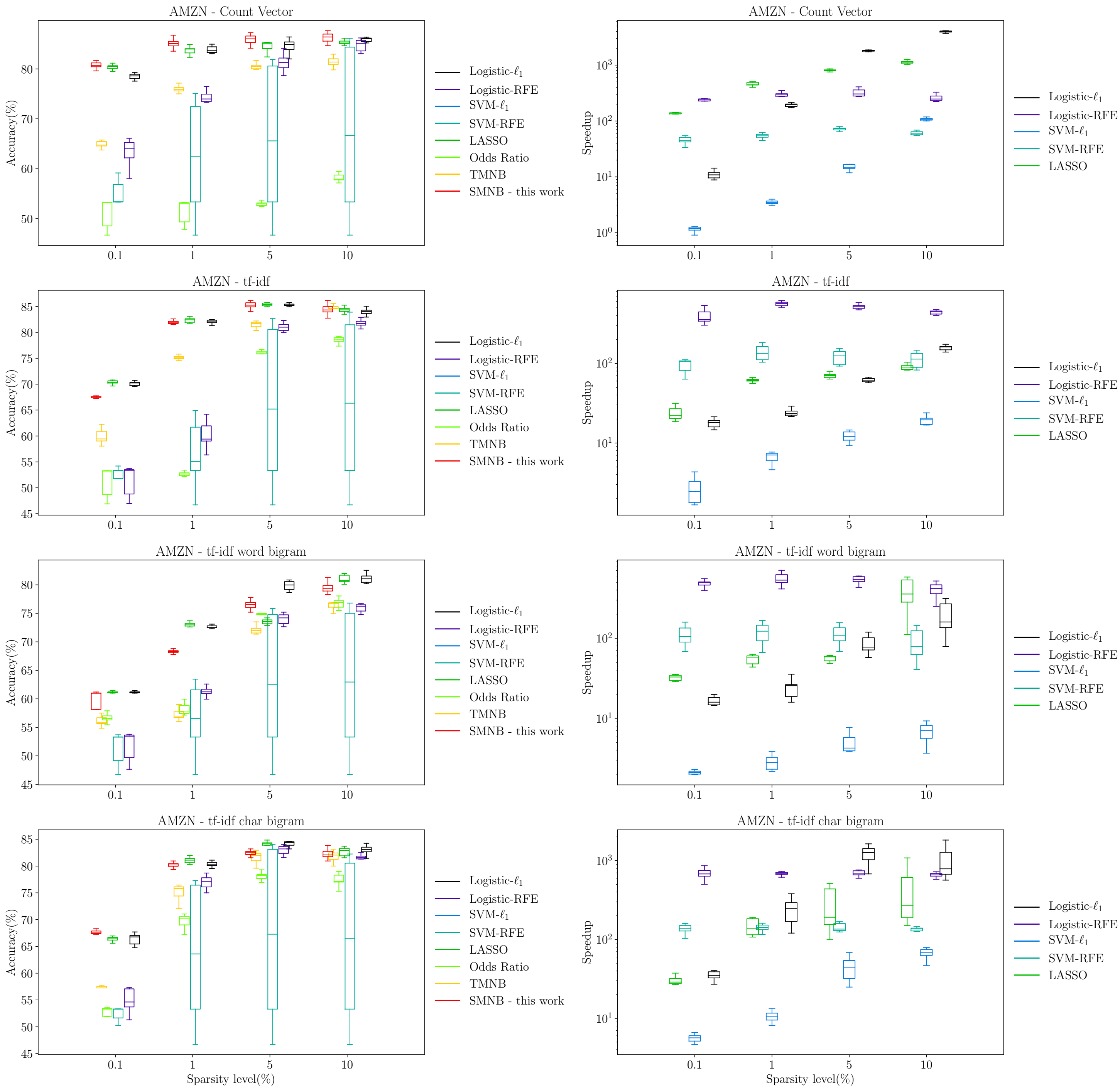

Figure 4: Experiment 2: AMZN - Stage 2 Logistic

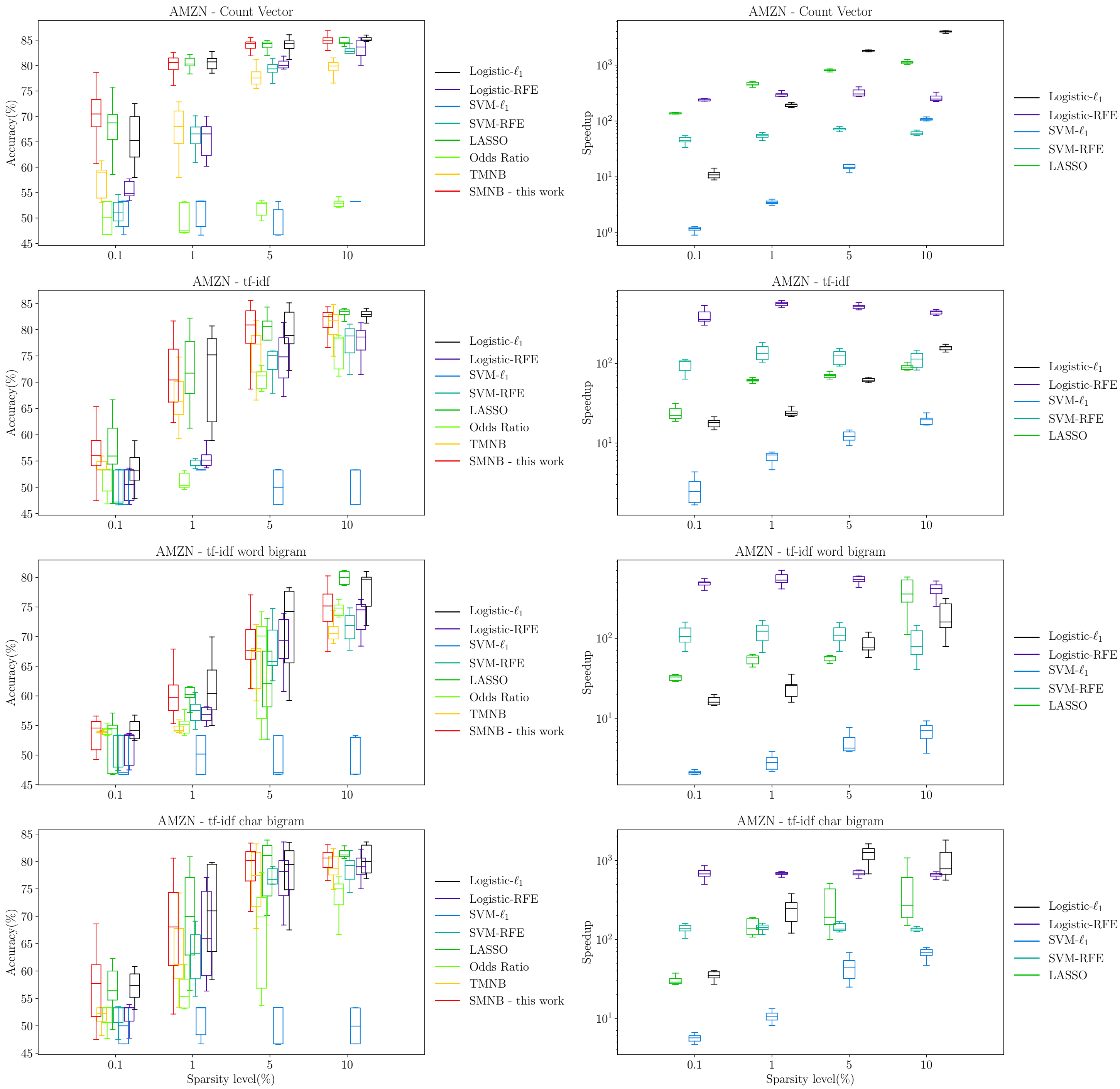

Figure 5: Experiment 2: AMZN - Stage 2 SVM

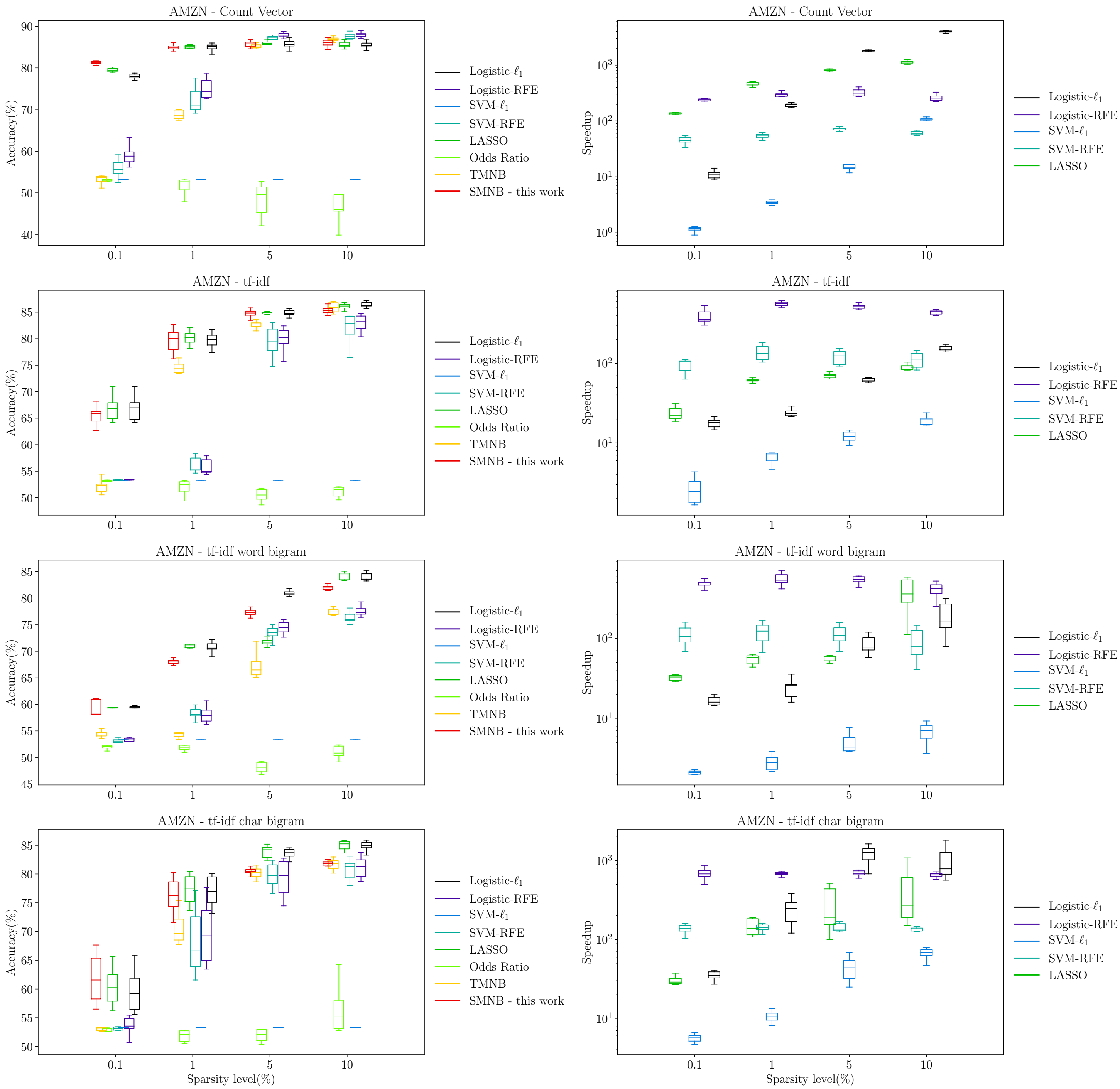

Figure 6: Experiment 2: AMZN - Stage 2 MNB

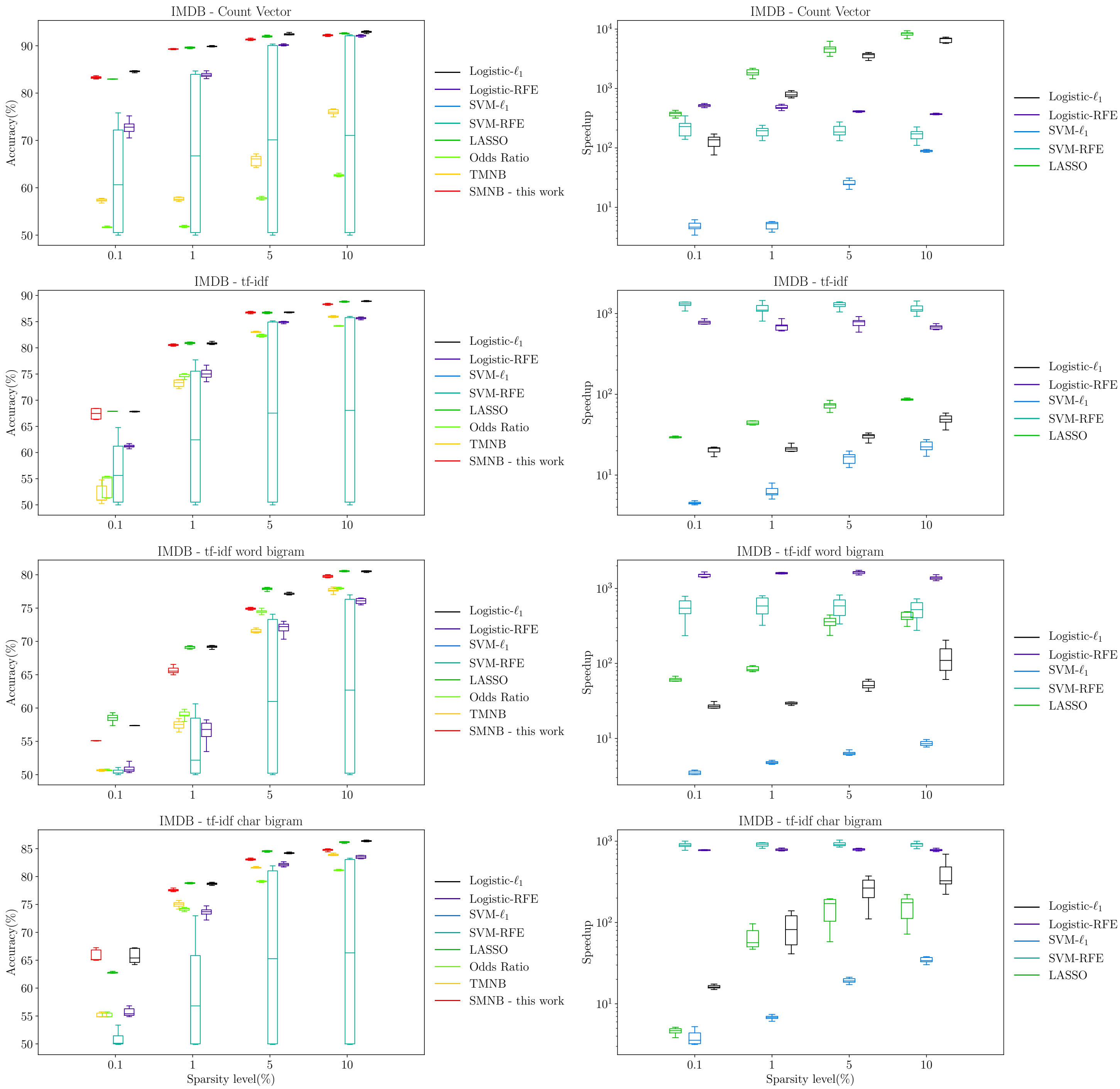

Figure 7: Experiment 2: IMDB - Stage 2 Logistic

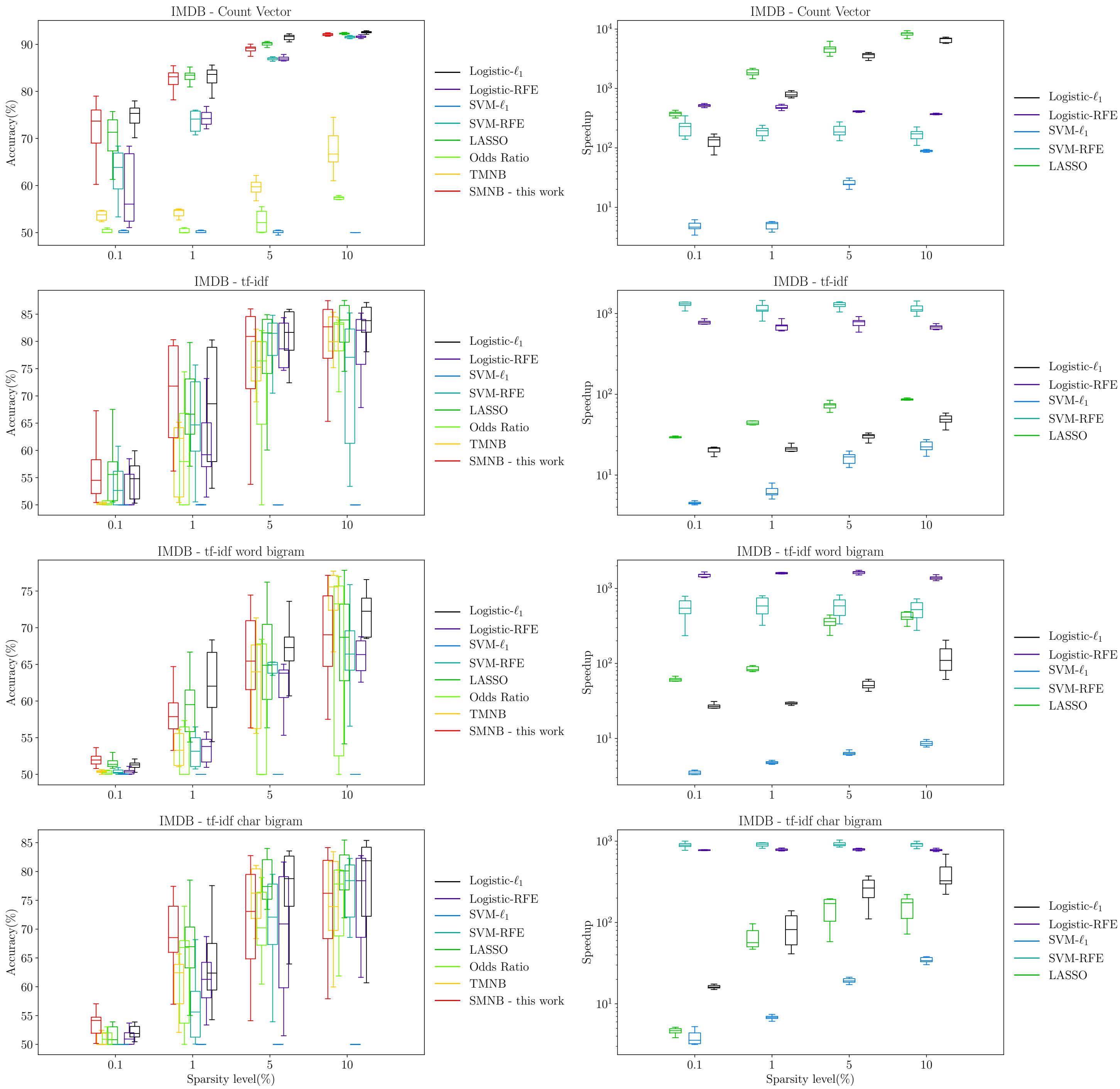

Figure 8: Experiment 2: IMDB - Stage 2 SVM

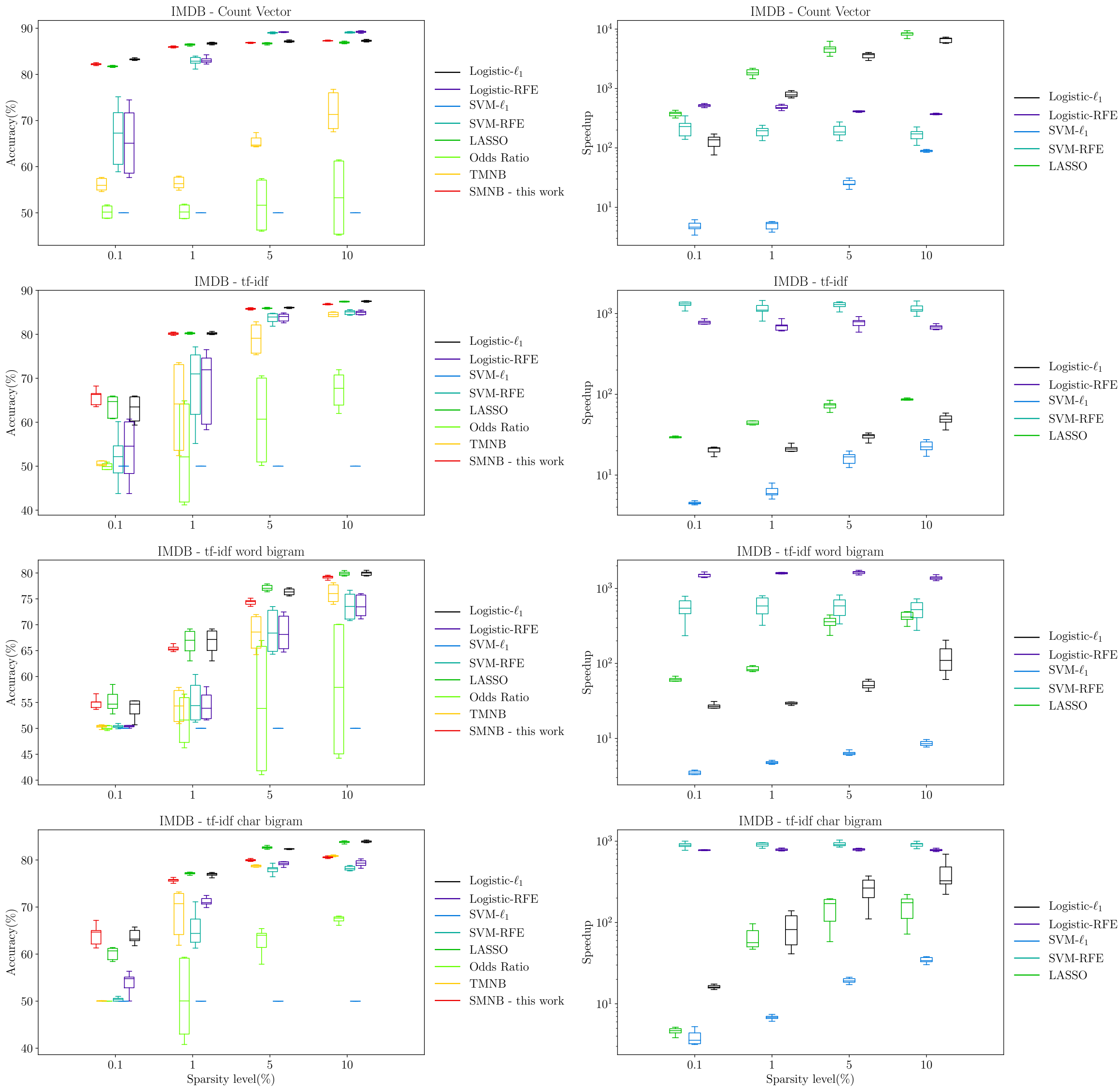

Figure 9: Experiment 2: IMDB - Stage 2 MNB

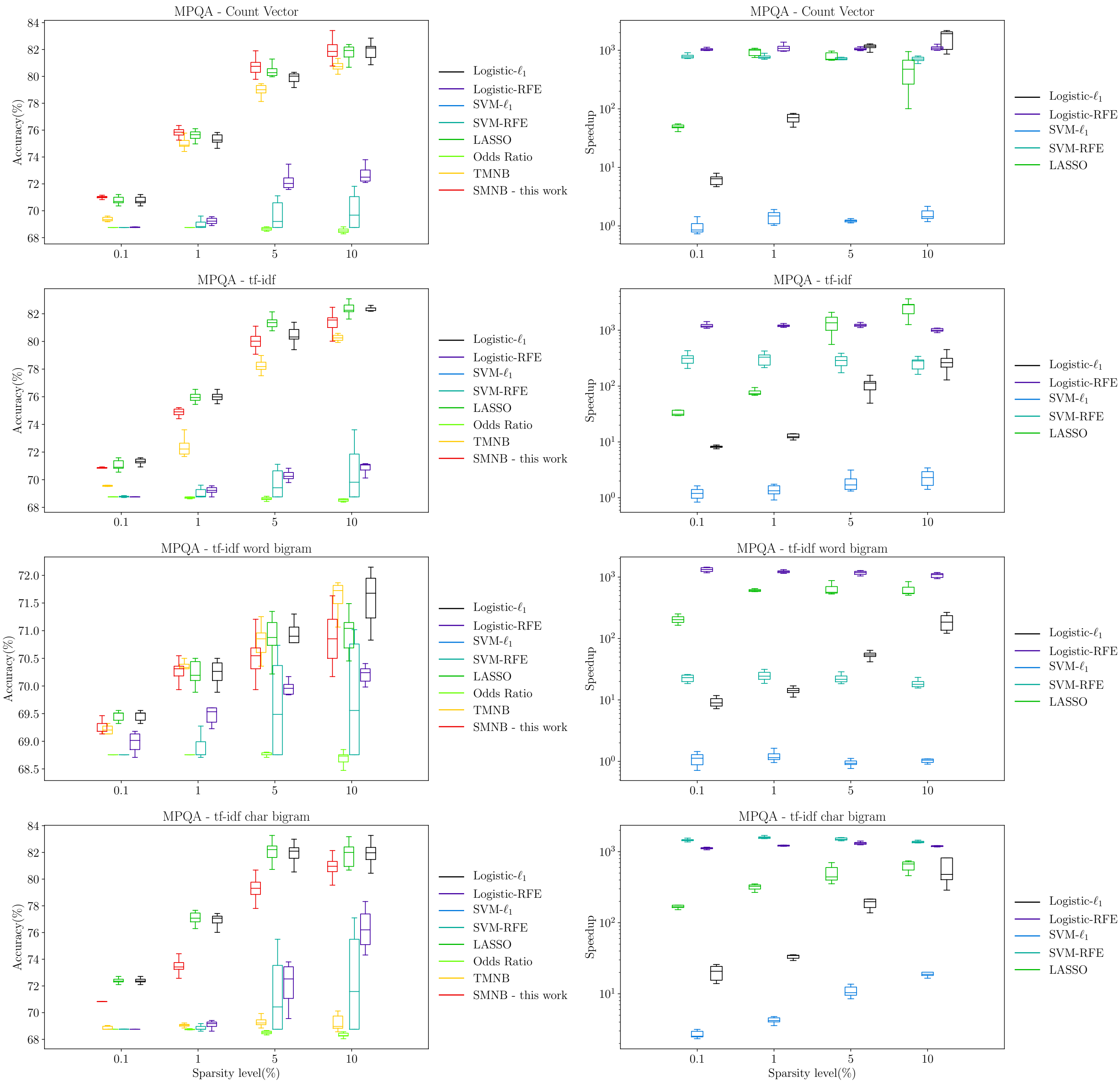

Figure 10: Experiment 2: MPQA - Stage 2 Logistic

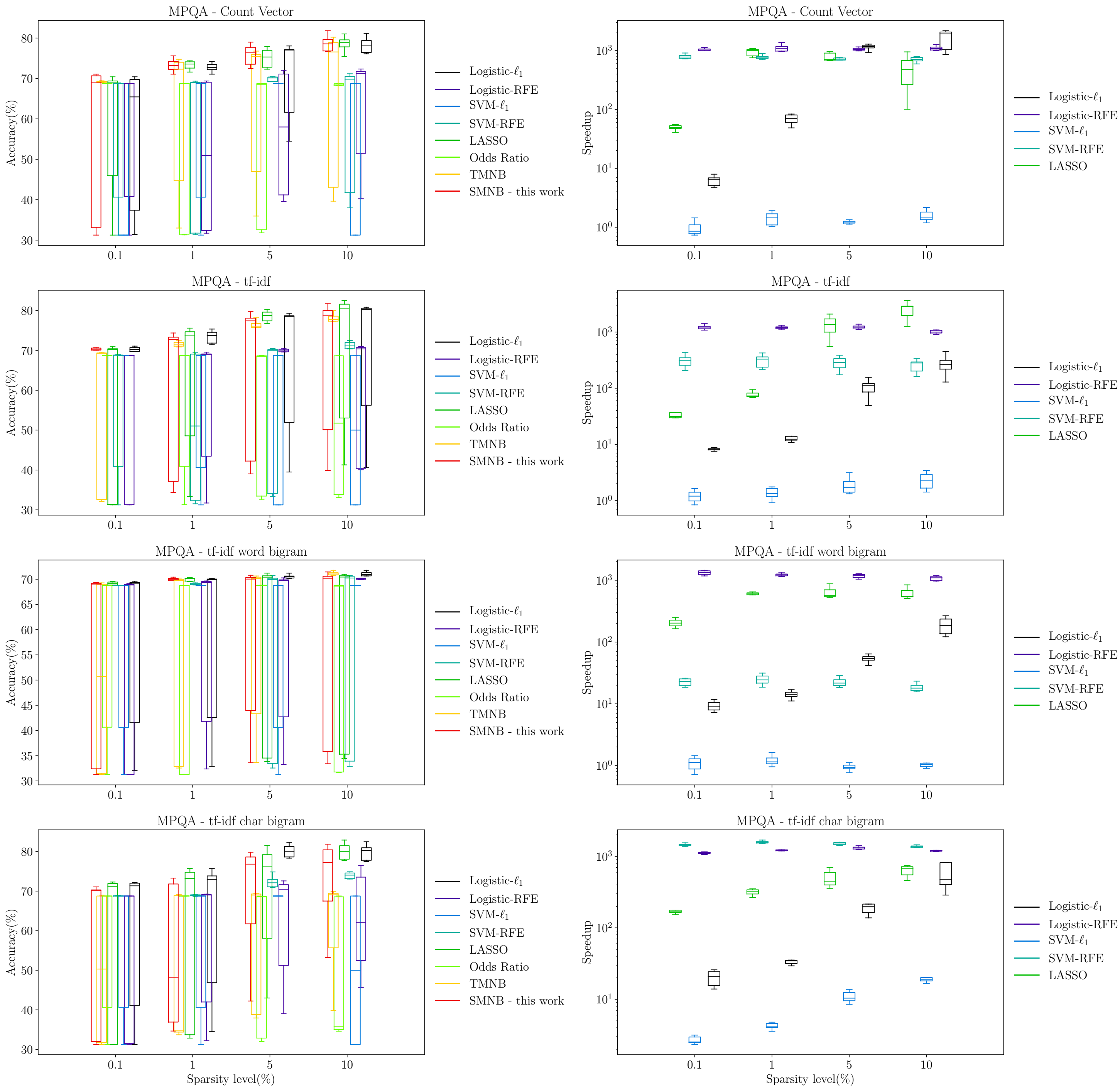

Figure 11: Experiment 2: MPQA - Stage 2 SVM

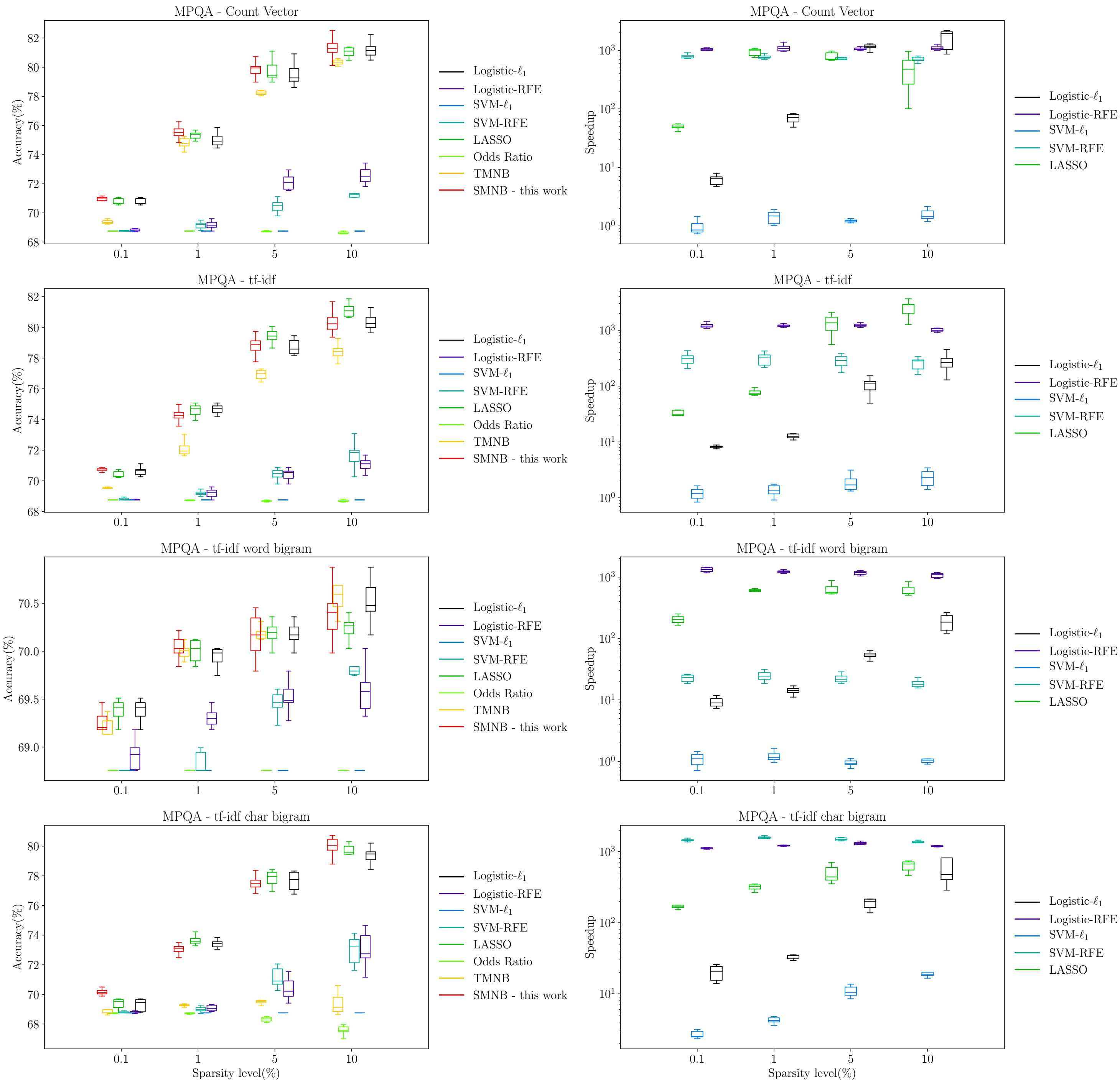

Figure 12: Experiment 2: MPQA - Stage 2 MNB

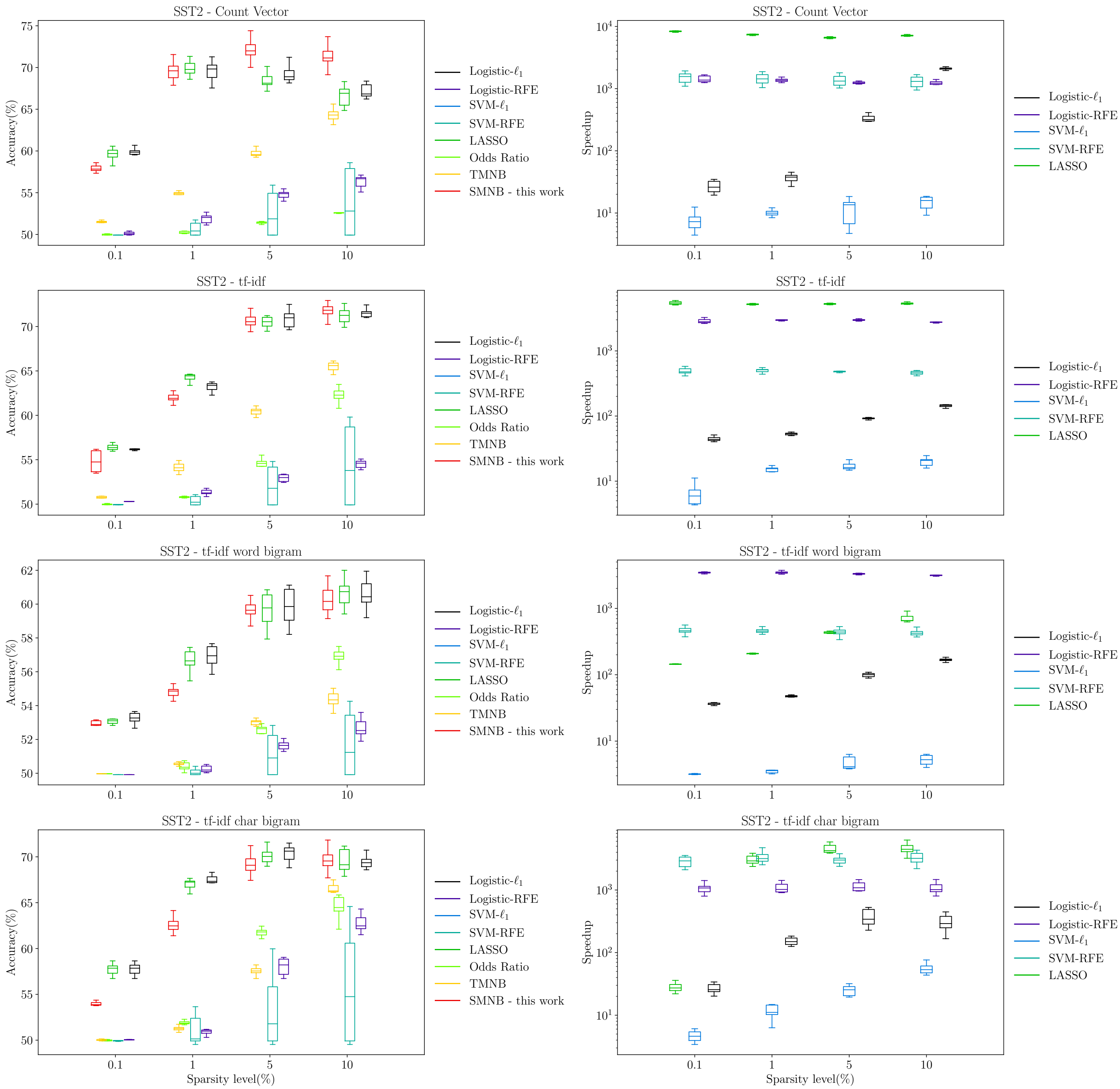

Figure 13: Experiment 2: SST2 - Stage 2 Logistic

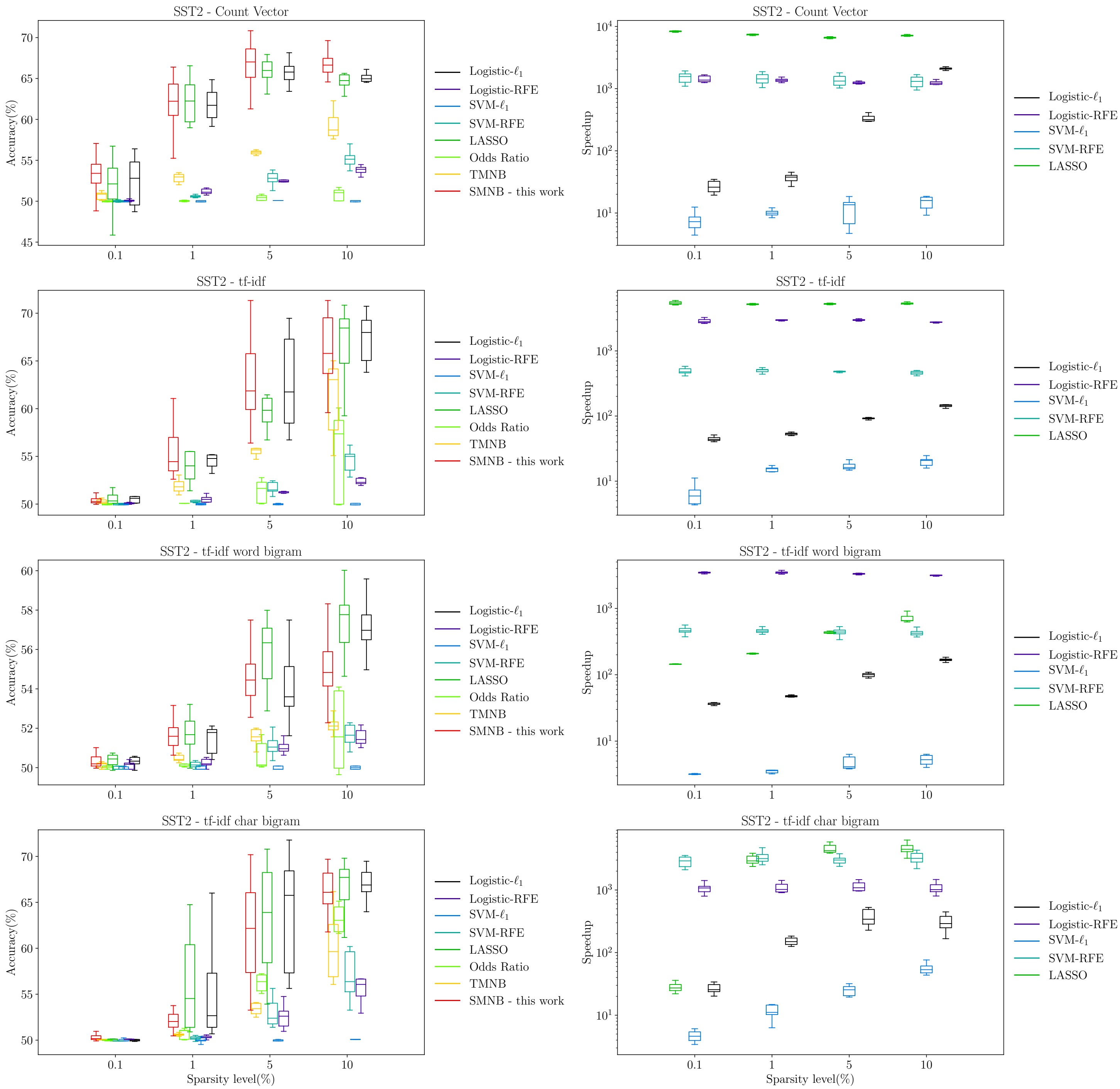

Figure 14: Experiment 2: SST2 - Stage 2 SVM

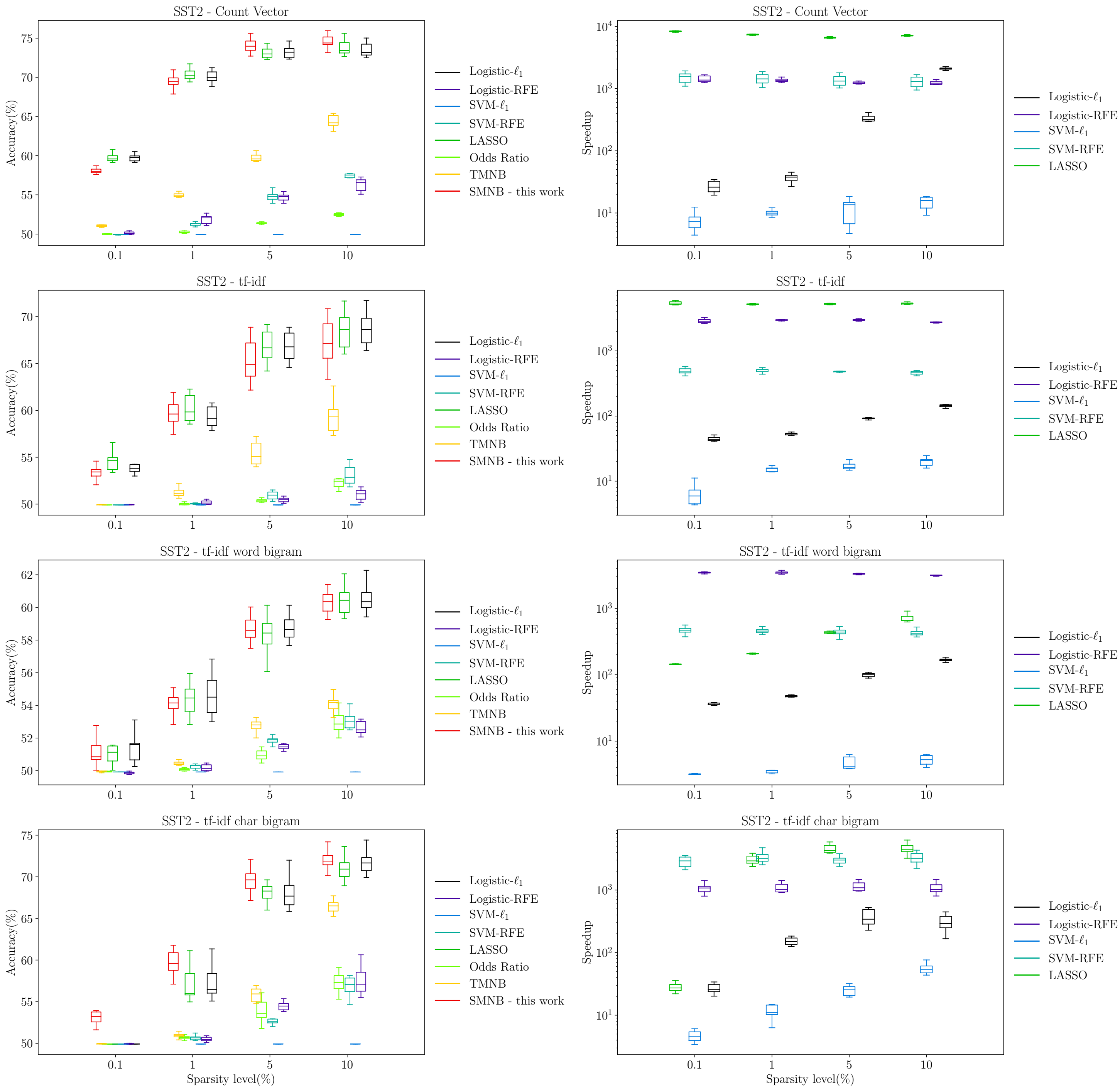

Figure 15: Experiment 2: SST2 - Stage 2 MNB

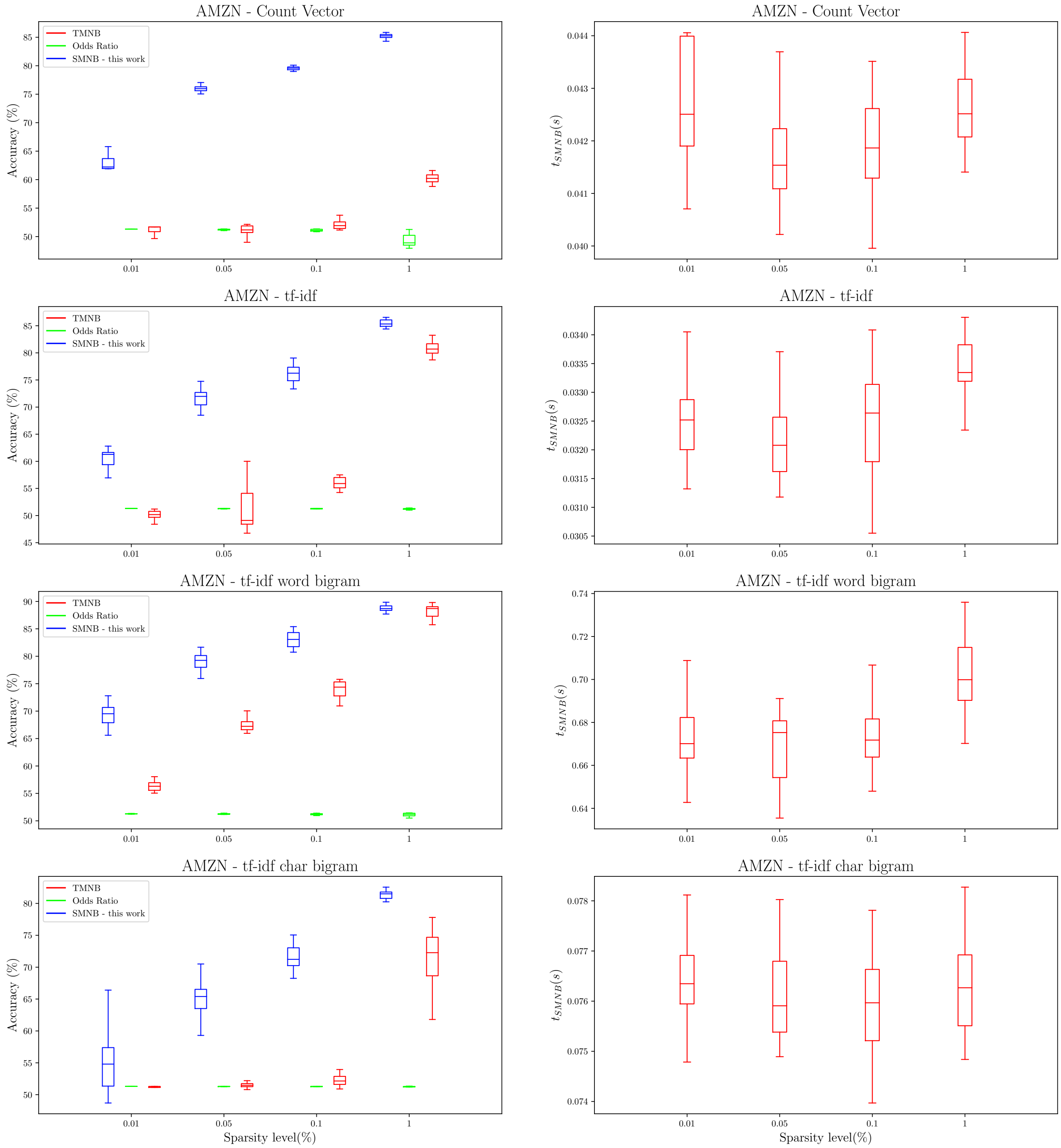

Figure 16: Experiment 3: AMZN - Stage 2 MNB

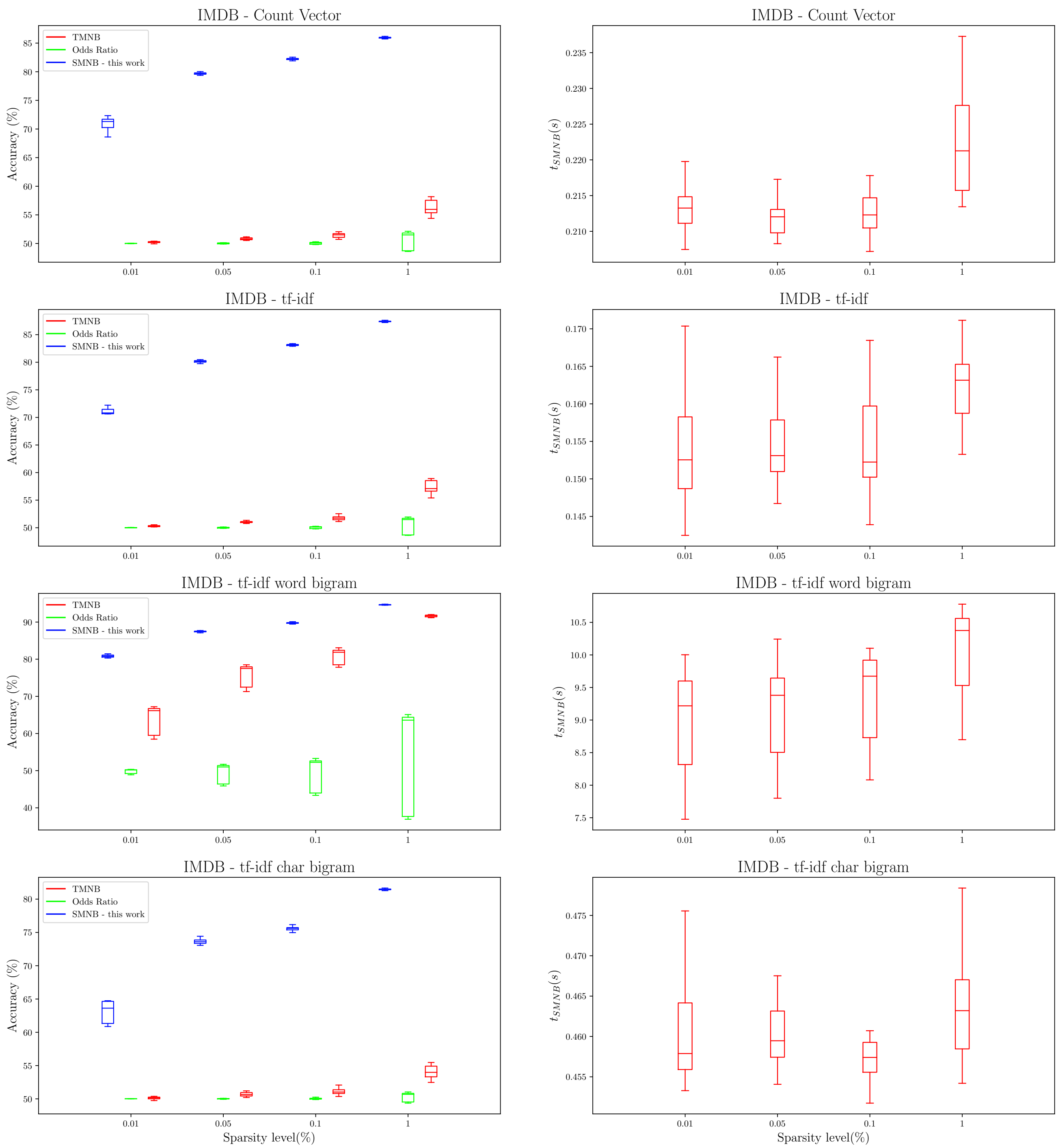

Figure 17: Experiment 3: IMDB - Stage 2 MNB

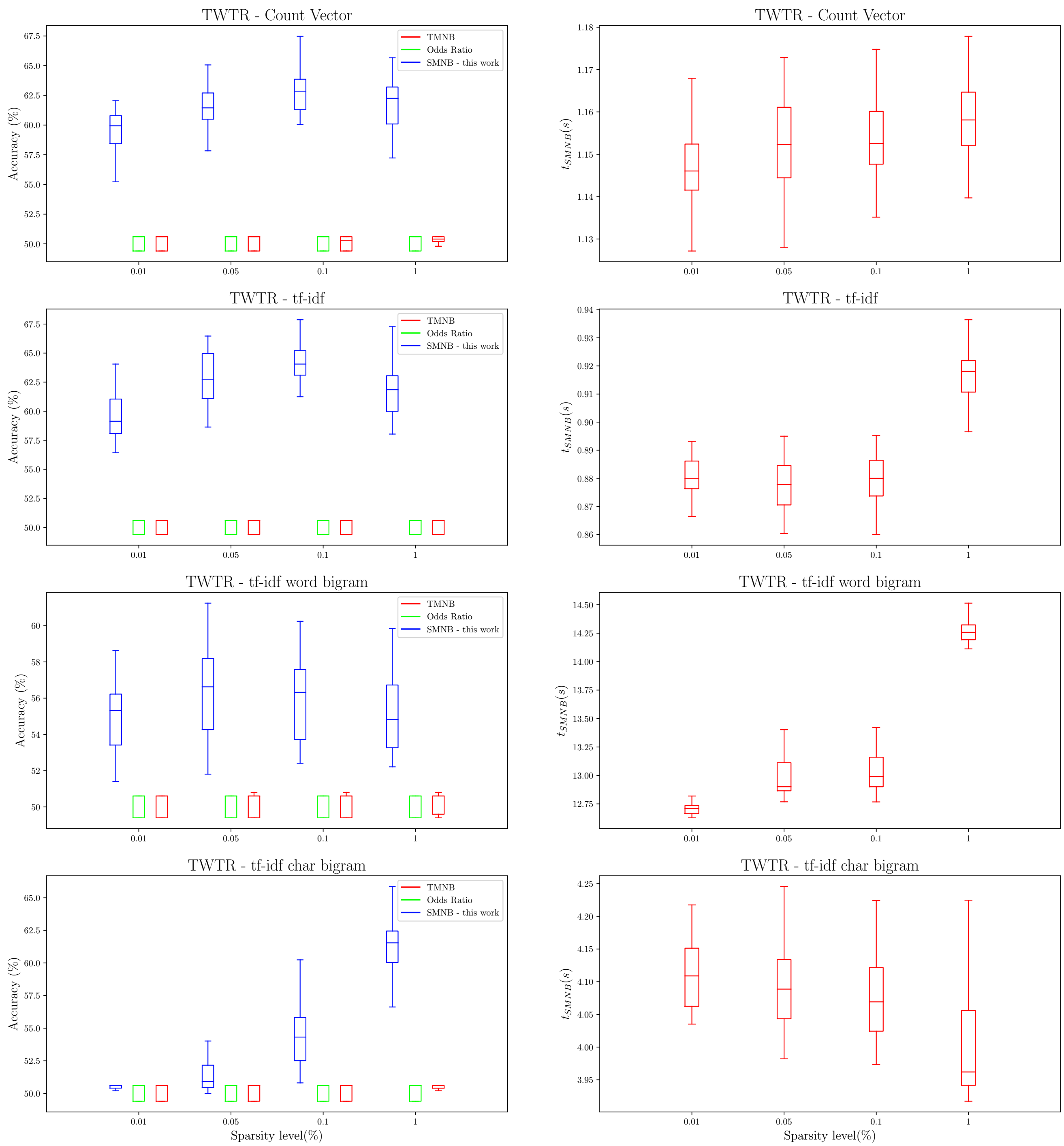

Figure 18: Experiment 3: TWTR - Stage 2 MNB

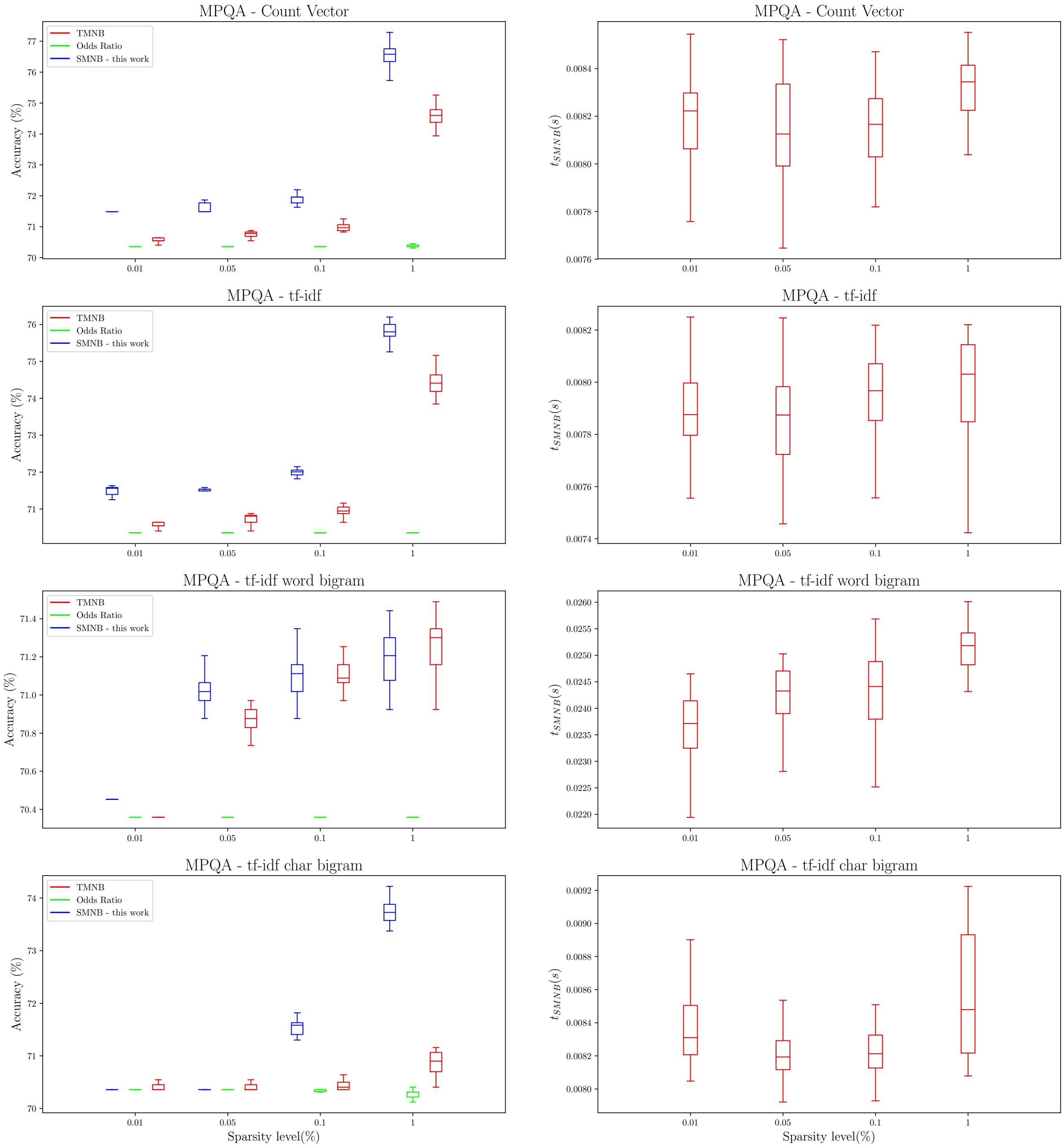

Figure 19: Experiment 3: MPQA - Stage 2 MNB

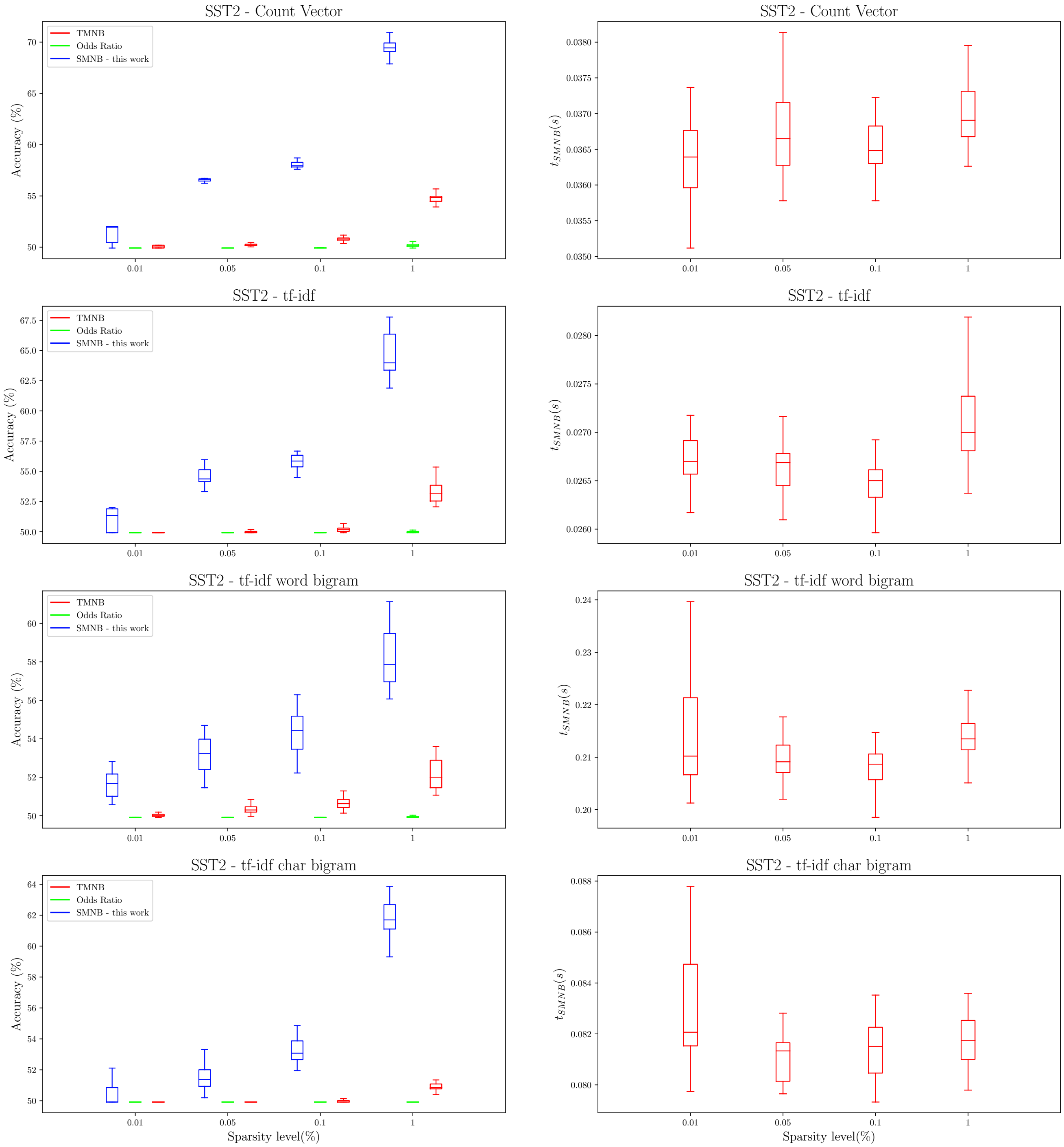

Figure 20: Experiment 3: SST2 - Stage 2 MNB